STUDYING RARE PATIENTS WITH COMMONLY-AVAILABLE INFORMATION: SOCIAL MEDIOMICS FOR RESEARCHING PATIENT HISTORIES IN AUTOIMMUNE HEPATITIS (AIH)

Anand Kulanthaivel

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics and Computing

Indiana University

December 2019

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Josette F. Jones, R.N., Ph.D., Chair

_____

Stasa Milojevic, M.L.I.S., Ph.D.

December 10, 2018

_____

Craig S. Lammert, M.D.

_____

David J. Wild, Ph.D.

ACKNOWLEDGMENTS

The author would like to acknowledge first and foremost to the studied Facebook Group and all of those who survive daily with autoimmune hepatitis (AIH), congenital disorders, and the aftermath of cancer: All of these are incurable but medically and socially neglected conditions.

The author also wishes to acknowledge the phenomenal amount of unconditional moral support given by his friends and mentors throughout the past two decades in achieving the difficult tasks of completing Bachelor and Master degrees, spending several years in the Biotechnology industry, and finally, as towards completing the Doctorate degree.

Finally, the author wishes to acknowledge those who supported this work academically, in particular his academic and research committees, including (Dr. J.F. Jones, IUPUI SOIC) and minor (Dr. S. Milojevic, IUB SICE) advisors. The author is furthermore indebted to J. Patel (IUPUI SOIC) as a co-annotator, colleague, and friend.

Finally, but not at the least, it is well-known that mentorship is a two-way street: The author wants to acknowledge all of his students and mentees, including his Health Information Management students and Health Informatics Master-level mentees.

Anand Kulanthaivel

STUDYING RARE PATIENTS WITH COMMONLY-AVAILABLE INFORMATION: SOCIAL MEDIOMICS FOR RESEARCHING PATIENT HISTORIES IN AUTOIMMUNE HEPATITIS (AIH)

Autoimmune Hepatitis (AIH), an incurable chronic condition of unknown cause where the body attacks its own liver, is a rare disease, with a current diagnosed worldwide prevalence of < 150,000. Inadequately treated, AIH can cause progressive liver damage and ultimately liver failure. A wide variety of symptoms are associated with AIH including severe fatigue, joint pain, depression, anxiety, and insomnia.

While precision medicine's genomics has attempted to shed light on the disease, other non-molecular "-omics" approaches can be taken in studying AIH patients, who often utilize social media to gather information from other patients or care providers to apply to their own AIH disease course. It is proposed that these patient-generated social mediomes can create self-report health records for patients – and facets of their lives - otherwise unreachable by conventional research.

In this feasibility study, I examined in an exploratory fashion the social mediome of a large (N > 1000) gathering of AIH patients and caregivers as present on a Facebook Group to determine the potential of mining various types health-related user information. The following types of information were mined, with *feasible* indicating a reliability of F >= 0.670:

1) Types of health information shared and structures of information sharing (Feasible)
2) Types and directionality of support provided by and to users (Portions feasible)

3) Clinical factors (AIH-related and otherwise) disclosed by users
    a. Medication intake (Feasible)
    b. Signs and symptoms (including pain and injury) and diagnosed comorbidities (Portions feasible)
    c. Results of disease monitoring blood tests (Portions feasible)
4) Contextual (non-clinical; environmental; social) factors disclosed by users (Detection of which type of factor discussed occasionally feasible).

The resulting knowledge is required to adequately describe the disease not only clinically, but also environmentally and socially, and will form part of the basis for future disease studies.

Josette F. Jones, R.N., Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF DEFINITIONS & ABBREVIATIONS

- **AIH (Autoimmune hepatitis):** a disease where the body's own immune system attacks the liver. Unrelated to viral hepatitis disorders. Occasionally known as *Autoimmune Liver Disease (ALD; AILD), Hepatic Lupus, or Liver Lupus.*
- **AC:** Stands for "administering colleague" and refers to a hepatologist colleague of the author's, who administers the researched Facebook™ group.
- **Annotation (also "annotating", "annotated", etc.):** The procedure of a human reading a document and marking features about it.
- **Autoimmune:** Indicating any disorder where the body's own immune system attacks other parts of the body.
- **Comment:** On Facebook™, refers to a reply made to a post.
- **Corticosteroid(s):** A class of immunosuppressive medications related to cortisol, a natural human hormone.
- **Emotional Support:** A type of verbal support given to enhance the receiver's emotions. Used here as well as a synonym for *social support*.
- **Etiology:** The root cause(s) of a condition
- **Facebook™:** The world's largest online social media (SM) venue.
- **FML:** In this dissertation used to refer to **Facebook Metalayer Language**, a novel format used to store data and metadata about user-generated content from Facebook™.
- **Genotype:** A genetic characteristic of a person, determined by DNA sequence.
- **Hepatic:** Refers to things that have to do with the liver
- **Immunosuppressant (also "immune suppressant" or "immunosuppressive"):** Refers to a drug that reduces the functioning of the immune system.

- **IPI (Identifiable Private Information):** referring to information that can be used to contact an individual. The exact categorization of what does and does not constitute IPI may vary by jurisdiction.
- **Mediate (also "mediated" or "mediates"):** To have a role in the occurrence of something. If *A mediates B,* this indicates that A has a role in B's occurrence.
- **Morbidity:** Suffering, both subjective and objective, caused by a condition
- **Mortality:** Death caused by a condition
- **NLP (Natural Language Processing):** referring to a variety of tools and algorithms that mine, process, and classify large bodies of text.
- **Pathogenesis:** The way by which a disease starts.
- **Phenotype:** Something about a person or organism that is not described by a gene sequence (common examples include hair color and laboratory test result values).
- **Post:** On Facebook™, refers to a top-level posting. Comments are replies made to these posts.
- **Rare disease(s): C**onditions that each affect fewer than 1 in 1,500 Americans but together affect over 10%. Also known in some parlances as *rare disorder(s)* and/or *rare condition(s).*
- **SM (Online social media):** a diverse set of Internet-based sites and services that offer platforms for anybody to communicate and publicize. **Venue** is a suffix that refers to a specific instance of online social media.
- **User-generated content:** Content (text postings) made by everyday users over online social media (SM). Similar to patient-generated data, except the users are not necessarily patients and what they post is usually free text.
- **User narrative:** In this dissertation, refers to the entire communication text of a single social media venue user over that venue, compiled into a single, human-readable file.

- **Xenobiotic:** Referring to any substance or chemical that is foreign to the human body. Most xenobiotics studied in medicine are pharmaceutical products (drugs), but other important classes of xenobiotics are over the counter medications and certain chemicals (vitamins, minerals, and various other organic small molecules) present in foods and herbal supplements.

CHAPTER 1. INTRODUCTION TO THE PROJECT

1.1 Summary Introduction

The increasing globalization and inter-connectedness of shared information has benefits and disadvantages. The inter-connectedness of information shared in the non-research realm, while leveraged by major corporations for profit, is insufficiently leveraged by the biomedical field for underserved communities, including those with conditions (i.e., rare diseases) that are neglected by the current research establishment.

Autoimmune Hepatitis (AIH) is an example of a rare disease that is underserved by current research: Different from the more commonly known viral hepatitis disorders, it is a chronic, immune-mediated, and incurable disease of the liver that creates significant morbidity in affected patients. Studying patients with this disease is difficult due to the fact that its rarity creates inherent geographic barriers. Furthermore, the etiology of AIH is multifactorial and includes variables that are beyond reach of detection by clinicians during office visits.

Online social media (SM) consists of a set of Internet-based venues that allow for peer-to-peer and global sharing of communications; SM has a very high global rate of usage. SM has been shown to be a productive and non-invasive venue for gathering patient-generated data pertaining to disorders both rare as well as common. Particularly prevalent are online SM venues that cater to specific patient disease groups.

It is therefore hypothesized that histories of patient clinical variables, including medication usage, AIH-related symptoms, self-reported clinical comorbidity, and clinically relevant environmental factors can be discovered via reading the posts of affected individuals over social media. Due to the importance of the provision of social support in the care of rare diseases, it is also important to study how social support is offered to and between those affected by AIH and the effects of the social support offered.

Most importantly, a colleague of the author's, know here as AC (Administering colleague), administers and moderates a Facebook™ group online social venue[1] for patients with AIH. Because online SM venues are frequently used to share condition-related information, the types of AIH-related information shared between patients, caregivers, researchers, and providers can be assessed via analysis of this Facebook™ group.

1.2 Goal & Purpose

Thus, by utilizing patient-generated data from an AIH-specific SM venue, the overarching goal is to assess if social media communications can be analyzed to enhance the body of knowledge that surrounds this disorder, and therefore create a knowledge framework that may one day be used to improve clinical and everyday care of those with AIH.

1.3 Aims & Hypotheses

Specific Aim A: To identify and categorize the body of information that is distributed via a large AIH-related SM venue and communication structures by which this information is shared. The information that are shared over SM are diverse and may include personal patient stories, general advice from practitioners, research-related information, and health-related advertising. Describing the types of information found as well as finding out how it is shared is expected to help enhance future patient disease resources.

---

[1] *Note:* The identity of this Facebook™ group is withheld from this dissertation for reasons of privacy.

Specific Aim(s) B1/B2/B3: To determine the feasibility of analyzing AIH-related social media content from an AIH-related Facebook group better inform the clinical and social research community about the disorder, enhancing knowledge of:

- Sub-Aim B1: Pharmaceutical treatments discussed by AIH patients
- Sub-Aim B2: Clinical symptoms and comorbidities that impact quality of life of AIH patients
- Sub-Aim B3: Contextual and environmental factors that could potentially correlate with the AIH disease process

Specific Aim C: To determine the structures and types of social support offered to and between AIH patients and caregivers over SM communications on an AIH-associated SM venue. Due to the importance of social support provision in the care of incurable diseases such as AIH, it is required to study the types of and quantity of social support that is offered to and between AIH affected individuals over SM communications.

The hypotheses of this dissertation's research are in a one-to-one relationship with the specific aims. Because the research conducted in this dissertation is of the exploratory proof-of-concept nature, hypotheses are all alternate and center on feasibility goals and not explanatory ones.

Specific Aim A Alternate Hypothesis: The health-related information communicated between patients, caregivers, clinicians, and researchers over an AIH-associated SM venue can be analyzed in order to determine its origin and subject matter.

Specific Aims B 1/2/3 Alternate Hypotheses: The alternate hypotheses for each sub-aim under Specific Aim B are similar; they all state that a certain type of factor (pharmaceutical product

usage, symptoms, or contextual/environmental factors) can be discovered through studying social media communications of users in an AIH-associated SM venue.

Specific Aim C Alternate Hypothesis: Types and contents of, and communication metrics of, social support given between AIH patients and caregivers can be detected via analysis of an AIH-associated SM venue.

CHAPTER 2. GENERAL BACKGROUND & LITERATURE REVIEW

2.1 Rare Diseases: An Introduction

As defined by the National Organization for Rare Disorders (NORD), rare diseases comprise conditions that each affect fewer than 1 in 1,500 Americans (i.e., fewer than 200,000 individuals across the US population) [1]. Over 7,000 rare diseases are recognized by NORD and the total prevalence of diagnosed Rare diseases is estimated to be approximately 10%.[2] Rare diseases are, for the most part, not curable. This problem, combined with the fact that Rare diseases may present extreme morbidity and often premature mortality to patients, makes rare diseases a significant nationwide burden on quality and quantity of life.

The situation of rare disease patients is exacerbated due to the fact that although current research in Rare diseases is progressing well, it still lacks optimal quality and quantity. Writing for *Clinical Pharmacology & Therapeutics,* Smith (2016) states that rare disease is a topic where "there are no experts among us" (p. 312)[3]. The inherent rarity of these conditions creates major barriers to recruiting and retaining patients in research and treatment. It is also noted that rare disease patients themselves have geographical disadvantages: rare disease patients also often travel large geographic distances in order to participate in research and even to seek treatment. [4, 5]

2.2 Autoimmune Hepatitis (AIH): Epidemiological, Etiological, Clinical

Next, it is elaborated the clinical entity that is the object of this dissertation: Autoimmune Hepatitis (AIH), being a rare disease, inherits the research and treatment issues seen with other rare diseases. Its extreme rarity (less than 1.2 out of 100,000 in the US are diagnosed[6]) and exceptional burden of morbidity create further significant difficulties in studying populations with the disease and can also make treatment and daily life with the disease very difficult for patients.

Differences in researching adequate sample sizes of AIH patients are exacerbated by its multifactorial etiology.

One factor proposed to be involved in the etiology is viral disease. Infectious diseases that may predispose individuals to AIH include prior infections including Epstein-Barr Virus (EBV) and Hepatitis C Virus (HCV), but It has notably been observed that correlation of AIH to EBV and HCV infection is difficult due to the small sample size of available AIH patients for direct study.[7] Similarly, one case report series discussed a potential association of AIH with human immunodeficiency virus (HIV), but with only three case studies.[8]

Non-AIH autoimmune disorders are widely thought to be found at a much higher rate in AIH patients than in the general population. The disorder most commonly thought to be comorbid with AIH is autoimmune thyroiditis (also known as Hashimoto thyroiditis), considered comorbid by Wong & Heneghan's review with between 8% and 20% of all AIH cases.[9] Rheumatoid arthritis (RA) was also noted by this review to be comorbid, with an estimated prevalence of 4-6% in the AIH population. The association between AIH and celiac disease (a form of autoimmune-mediated gluten intolerance) was demonstrated in one case study.[10] Furthermore, researchers claimed the development of AIH in a patient with psoriasis not controlled by immunosuppressants.[11]

Genetic correlates exist for AIH, but the literature evidence from the author's perspective appears to still be sparse. One study showed that alpha1-antitrypsin (A1A) deficiency phenotypes were more prevalent in patients diagnosed with Type 1 AIH (a subtype of AIH).[12] The expression of TIPE2, a tumor necrosis factor-induced protein, was found to be altered in a mouse model of autoimmune hepatitis, but as of that study (March 2017), no human correlation had been reported. [13] In humans, genetic culprits invoked by researchers include certain human leukocyte antigen (HLA) subtypes, with several anomalous variants of these genes being modestly correlated in one genome-wide association study with the presence of AIH.[14] Therefore, studying AIH beyond the patient genome is desired.

Xenobiotics, a collective group of chemicals encompassing most pharmaceutical products, many nutritional supplements, and most environmental toxins, are defined as compounds or substances that are foreign nature to the body.[15] Researchers have positively correlated intake of various xenobiotics with the development of AIH. Antibiotics (in particular, preparation combinations thereof) are perhaps the largest treatment group of medications known to correlate with AIH development. The antibiotic combination preparation trimethoprim-sulfamethoxazole (TMP-SMZ; TMP-SMX) has also been associated with the development of AIH.[16] Another common antibiotic combination, amoxicillin with clavulanic acid (AUGMENTIN; Amox-Clav; Amoxicillin-Clavulanate) has been hypothesized to increase the chances of developing AIH. Similarly, tetracycline antibiotics have been implicated in the pathogenesis of drug-induced AIH in a series of case studies and also in single, modestly sized AIH cohort. [17]

Other medication families and medications have been suggested by researchers as AIH *triggers*. Case reports long implicated the statin family of drugs in the pathogenesis of AIH.[16, 18-20] In addition, use of the drug nitrofurantoin was found to precede a form of drug-induced liver injury with microscopic features of AIH.[21] A separate review of hepatotoxic medications also implicated the pharmaceutical products carbamazepine, chlorpromazine, methotrexate (a treatment option for rheumatoid arthritis, commonly comorbid with AIH), ibuprofen, and valproates.[22]

Variables influencing and confounding the etiology of AIH are hypothesized to reach beyond what is detectable in the clinic and extend into the patient's environment; these variables often are also associated with xenobiotic exposure. The common herbal supplement black cohosh has been implicated in the pathogenesis of AIH.[23] Another environmental variable is hypothesized to be exposure to environmental pollution; trichloroethylene (TCE), present in low levels in much of the United States' water supply, has been implicated in autoimmune liver

diseases including AIH [24], and the organic compound vinyl chloride has been suggested to increase the risk of liver lesions (including those that are autoimmune-mediated). [25]

Diet, an important part of the patient environment, may also have a role in the pathogenesis of AIH. Due to the aforementioned finding of celiac disease association with AIH, it is hypothesized that gluten consumption could be an etiological or aggravating factor in some patients with AIH.[10] A population-based study in New Zealand found that positive past histories of alcohol usage, despite that chemical's well-known hepatotoxic effects, were correlated with a lower prevalence of AIH. [16]

In summary, the diversity of etiological variables – particularly those seen outside of genetics -- correlated to AIH makes it significant to study every potential factor that surrounds the AIH group user; clinically-oriented data such as drugs, signs, and symptoms are unlikely to divulge a full picture of the AIH group user. Furthermore, the clinical environment usually cannot ascertain environmental factors, making the study via social media of AIH patients' day to day lives of potential benefit.

The possibility exists that xenobiotics other than those currently implicated may still be implicated in AIH's pathogenesis. General liver toxicity due to medications is known as drug-induced liver injury (DILI), but the International Autoimmune Hepatitis Group (IAIHG) states that DILI must be excluded from the diagnosis of AIH and yet simultaneously state that there exists no means by which DILI should be excluded.[26, 27] Furthermore, the concept of drug-induced autoimmune liver disease (DIAILD) does exist; it biochemically and clinically bridges AIH (and similar disorders) with DILI. The main anomaly in the relationship is that some forms of DIAILD are acute and resolve with drug discontinuation. Nonetheless, checkpoints of molecular cascades are shared in common between DILI, DIAILD, and AIH[28] Therefore, it is of interest to study the use of a general corpus of xenobiotics as potential triggers of AIH.

In addition, treatments (such those for AIH) that have a high side effect burden may in turn incur an even higher burden of xenobiotic load by convincing the patient, provider, or both

that more medications and/or supplements are needed.[29] Therefore, the potential for both beneficial and adverse polypharmacy may be studied by using a wide net across all potentially consumed xenobiotics.

Other substances whose use is not easily ascertainable in the clinic are also known to participate in AIH's etiology. The common herbal supplement black cohosh has been implicated in the pathogenesis of AIH.[23] Another published compendium[22] implies a larger list of supplements that may be responsible for DILI and DIAILD types of liver damage; the list includes not just black cohosh but also common supplements such as ginseng, ephedra, chamomile, milk thistle, *kava kava*, and pennyroyal.

Although medications (xenobiotics) are correlated with AIH etiology, more obvious correlates in the usage of xenobiotic pharmaceutical products by AIH patients exist as the disorder requires medical treatment with regimens of immunosuppressants. Pharmaceutical products commonly prescribed for AIH treatment include azathioprine, mechanistic Target of Rapamycin (mTOR) inhibitors, ursodiol (ursodeoxycholic acid), mycophenolate, and a variety of corticosteroids (chiefly prednisone, budesonide, and prednisolone). The use of cyclosporine is also common after liver transplant, a procedure required in the most severe cases of AIH.[30]

Comorbidities of AIH include recognized hepatobiliary sequelae as well as frequently co-occurring autoimmune diseases. Well-recognized hepatobiliary sequelae occur in autoimmune hepatitis (AIH) and, secondary to hepatic destruction, are comparable to those found in other hepatic diseases. Widespread and greatly feared is cirrhosis, a permanent gross scarring and damage of liver tissue. [31] Hepatic encephalopathy (HE), a form of dementia secondary to the liver's failure to properly filter out neurotoxic metabolites, may be found in more severe AIH cases.[32] Certain hepatobiliary tract autoimmune comorbidities, whose presence are frequent enough to have titular AIH-*overlap* syndrome designations[33], include the rare conditions primary sclerosing cholangitis (PSC) and primary biliary cirrhosis (PBC)[34], the latter of which significantly increases the risk of cirrhosis in AIH.[33]

Autoimmune comorbidities outside the hepatobiliary tract are also noted. A more recent and exceptionally wide-ranging study by Baven-Pronk et al (2018) discovered multiple comorbidities, with at least 2 (N > 150) patients each reporting autoimmune thyroiditis, celiac disease, scleroderma (systemic sclerosis), ulcerative colitis, diabetes mellitus, Crohn disease, and systemic lupus erythematosus (SLE).[35] The latter finding is unsurprising knowing that the anti-nuclear antibody (ANA) is noted to be one of those that may attack the liver in AIH. Similarly, Karp et al (2010) assembled a sizable electronic medical record (EMR) cohort involving over 200 AIH patients and found a strong association between AIH and Sjörgen Syndrome, an oral autoimmune condition.[36]

The disorder most commonly thought to be comorbid by Wong & Heneghan's review is in fact autoimmune thyroiditis (also known as Hashimoto thyroiditis or Hashimoto hypothyroidism), occurring in this opinion between 8% and 20% of all AIH cases.[9] An association between AIH and celiac disease (a form of autoimmune-mediated gluten intolerance) was demonstrated in one case study. [10] Furthermore, the development of AIH has been documented in a patient with psoriasis not controlled by immunosuppressants. [11]

Comorbidities with no direct known relationship to AIH are also observed. Physical pain is likely to be significant (see below section on physical pain in AIH), and the exceptionally high risk of disease-related psychological trauma may lead to increased experiences of reactive anxiety and depression in AIH patients. Diabetes Mellitus (DM) is a common disease that can affect the liver via fibrosis[37], steatosis[38], and generalized hepatic destruction.[39] AIH patients are equally prone to disorders that affect non-AIH-impacted patients, and it is uncertain how other disorders may influence the clinical course or quality of life of affected individuals.

Adverse drug effects (ADEs) are symptoms attributable by the clinician as secondary to medication intake. Typical ADEs are also known as *side effects* while less expected ADEs are known as *adverse drug reactions* (ADRs). Corticosteroids, a mainstay of AIH treatment, have a heavy burden of ADEs, some of which (such as edema, weight changes, and impaired WBC

function[40]) are common to original disease symptoms. Despite the beneficence of the treating clinicians, corticosteroids have been shown by at least one healthcare utility review to actually decrease quality of life of treated AIH patients compared to alternately-treated AIH patients.[41]

Immunosuppressive regimens with corticosteroids and other potent immunosuppressants such as azathioprine (AZA) also contribute to an increased risk of infectious diseases. In addition, such drugs and regimens are known to aggravate the paralyzing fatigue and psychological risks[42] already observed in untreated disease. As noted in the next section, pain is a common indirect sequela of AIH treatment.

Most importantly, this dissertation notes mentions of potential user ADEs (including pain ADEs; see below) as simply symptoms and not in a special ADE class. Specifically, the relationship will be viewed not in a clinical cause-and-effect knowledge framework, but instead in a scientific and associative one that is more suited the author's academic training expertise. Associative analyses[43] between drug intake (from the previous chapter) and signs and symptoms (from this chapter) can instead help elucidate (although not confirm) the relationships between drug intake and associated symptoms, which may or may not be ADEs.

In all chronic disease patients, pain (referring to physical pain in this article) is a significant negative impactor of quality of life. Pain has been hypothesized to be a significant ADE of AIH treatments. Although the reasons are poorly known, immunosuppressants such as mycophenolate[44,45] have been blamed as the cause for new-onset arthralgia in smaller studies. Indirectly, corticosteroids have been observed in non-AIH cohorts to increase pain burden by raising the risk of bone and joint degeneration, resulting in increased rates of painful injury including soft tissue injuries (bruises, sprains, strains, and similar)[42] and osteonecrosis-induced bone fractures.[46] The usage of some corticosteroids is strongly associated with other hepatobiliary derangements such as pancreatitis[47], which are often extremely painful.

Pain also occurs in AIH in manners that is currently not attributed to medication therapy. Likely due to hepatic inflammation and subsequent impingement on nearby sensory nerves, right

upper quadrant (RUQ) abdominal pain is frequently noted, and this pain may radiate to nearby regions of the abdomen. Autoimmune comorbidities of AIH include joint diseases[11] such as rheumatoid arthritis (RA) and psoriatic arthritis (PA), both of which feature arthralgia as a cardinal symptom.

Pain, referring to discomfort in the purest sense, is not just physical but also can be mental (psychological). Pain in the form of psychological distress (and potentially psychosomatic physical pain) is seen in AIH, which has neuropsychiatric disorder prevalence significantly higher than that of the general population and on par with infectious liver diseases (i.e., the viral hepatitides).[48] Mental depression is a significant QOL-limiting factor in AIH and other inflammatory liver diseases[49] and among the more feared of its neuropsychiatric sequelae.[50] Furthermore, immunosuppressant therapies (in particular, corticosteroids[51]) are associated with similar adverse neuropsychiatric effects.

Other psychologically-mediated inhibited states are noted. Fatigue (tiredness) is attributed to overall heavy illness burden but its specific etiology is still in the process of attribution.[52, 53] Fatigue is also commonly attributed as an ADE. [42] Finally, cognitive impairment is also noted as in AIH and can be a direct consequence of the sequela of hepatic encephalopathy,[32] and is also associated with depressed mood and fatigue.[54]

AIH-relevant disease characteristics range beyond symptoms and extend into objective laboratory signs, which are correlated with mortality and objective outward signs but less so with burdens of pain and fatigue. The monitoring of serum hepatic enzyme levels, conducted as the battery of liver function tests (LFTs), is the most significant cornerstone of AIH surveillance and management. Enzymes typically assayed in LFT batteries include alanine transaminase (ALT), aspartate transaminase (AST), alkaline phosphatase (ALP), and gamma-glutamyl transpeptidase (GGT).[55] In the setting of AIH, the lowering of these enzymes' levels close to normal baselines is the cardinal diagnostic measure for judging cycles of disorder remission and exacerbation.

Other hepatorenal parameters measured as surrogate remission measures include bilirubin and creatinine protein kinase (CPK). Another important object of monitoring are titers of autoantibodies, atypical proteins that form the basis for the body's autoimmune attack on itself. Seroprevalences of different antibodies, including anti-nuclear antibody (ANA), perinuclear anti-neutrophil cytoplasmic antibody (PANCA), and anti-liver kidney microsomal type 1 (anti-LKM-1) antibody are far in excess of those observed in the non-AIH population and in seropositive AIH patients are assayed as a surrogate for potential remission and exacerbation[56] even though seropositivity is not an explicit required criteria for AIH diagnosis.

2.3 Out of and Beyond the AIH Clinic: Contextual Factors

In disorders such as AIH where researchers believe to have weak clinical etiological correlates and protean signs and symptoms, the diametric opposite of biomarkers may be entertained; *patient contextual factors* are also investigated. Relating to the concepts of *environmental factors* and virtually a synonym to *social factors* and *social determinants of health (SDH),* contextual factors for purposes of this discussion are non-clinical factors present in any individual or population thereof. Commonly studied contextual factors include demographics (race/ethnicity, sex, gender, age), physical environment (geographic location environmental toxin exposures), economic factors (income, employment) and lifestyle-related factors (dietary behavior, use of recreational drugs, exercise/physical activity). Some contextual factors (in particular demographics) are easily and commonly ascertained at routine clinical visits, but a majority can only be elucidated with specific inquiry during studies or via observation of patients. Other contextual factors, often classified under *quality of life (QOL)* and *health-related quality of life (HRQOL),* focus on the patient's psychological and social well-being. Regardless of the type of contextual factor being investigated, the casual nature of social media allows for the potential of ascertaining non-clinical factors of a body of AIH patients.

In order to cast a broad net around potential contextual factors associated with AIH's risk profile, an exhaustive systematic review is performed on the literature regarding AIH risk factors in order to filter out the contextual risk factors that have thus far been elucidated by research.

The query (*AIH risk) OR (AIH factor)*[2] was over conducted over PubMed with a 10-year publication date filter with the intention of removing purely clinical articles from the search. 334 article abstracts were returned. Abstracts were then excluded by these criteria by the titles' indications, being assigned codes for reason of rejection:

- NHWRH: No focus on any liver disease at all (chiefly due to the authors using another meaning of the term *AIH*) (37/334; 11.1%)
- WRHEP: Sole focus on non-autoimmune hepatitides. (10/334; 3.0%)
- MCB: Only cell and molecular (including genomic and proteomic aspects) were being studied, and these aspects were not known to be environmentally influenced. (91/334; 27.3%)
- CLIN: Only clinical environmental factors (e.g., medication exposure, comorbid illnesses) were being studied (152/334; 45.5%)
- DEMOF: Contextual factor study in AIH, but only the study of demographic factors already ascertainable in the clinic (6/334; 1.8%)

After abstract filtering, 38/334 (11.4%) of articles remained; these were subject to full-text filtering by the same criteria as above. The following exclusions (in Table 1) were noted, with an additional criterion (ADVOC) to mark full texts that were not about contextual factors but advocated that contextual factors must be studied in addition.

[2] *Note:* As of the date it was searched, PubMed automatically expanded *AIH* into *AIH, autoimmune hepatitis, autoimmune liver disease, hepatic lupus,* and *liver lupus*.

Table 1. Exclusion of Articles Irrelevant to Contextual Factors

Excluded		Included	
CLIN	DEMOF	ADVOC	(Fully Relevant)
13/38 (34.2%)	4/38 (10.5%)	3/38 (7.9%)	18/38 (47.4%)

Synthesis

The remaining articles, in the categories Fully Relevant and ADVOC (21/38; 55.3%) are subject to a thematic synthesis. The themes to be entertained are diet-related factors, treatment noncompliance (patient-initiated withdrawal), physical environment, psychological quality of life, recreational substance usage, and research advocacy of more contextual factors research (ADVOC).

The most common theme was noted to be diet-related factors, defined as factors that are dependent upon the foods (and nutritional supplements) consumed by the patient (or other model organism). The interactions of the fatty liver syndromes non-alcoholic steatohepatitis (NASH) and non-alcoholic fatty liver disease (NAFLD) with AIH are documented. The interactions are important because poor diet (too high in caloric energy from fat and/or carbohydrates) is a prime suspect in the pathogenesis of these disorders.[38] Muller et al (2016)[57] demonstrated that NAFLD aggravated AIH in an anti-CYP2D6 murine (mouse) model. A study on humans (Himoto et al, 2017)[58] also suggested that fatty liver and AIH interact; this study found a deleterious effect on insulin resistance in patients with autoimmune liver features and NASH and also patients with combined autoimmune liver features and hepatitis C (HCV)-induced liver disease.

Outside NASH and NAFLD, epigenetic changes (changes in genetic expression) are hypothesized to affect the pathophysiology of AIH.[59] Although these changes are inherently genetic, they are byproducts of the environment directly modifying genetic expression: Common triggers for such changes include environmental exposure and SDH stressors, including but not

limited to dietary intake. The intestinal microbiome, heavily influenced by diet, may also mediate AIH's pathogenesis; Czaja (2017)[60] in fact recommends that dietary surveys of AIH patients be undertaken for this reason. As per Silva (2014),[61] adiponectin, a molecule synthesized by adipose (fat) tissue (which is increased by poor diet) has been correlated in increased levels with the development of hepatic cirrhosis, including AIH-related cirrhosis. Finally, a series of rodent models as reviewed by Luong et al (2013)[62] suggested links between dietary vitamin D consumption and a reduced risk of developing AIH-pathognomic histologic features.

Another salient theme observed was treatment noncompliance (also referred to here as *patient-initiated withdrawal*), defined for these purposes as a patient being unwilling or unable to take a medication required for AIH treatment. Van Gerven et al (2016)[63] note that many AIH patients become noncompliant with treatment due to the severe side effects of corticosteroids and other treatment immunosuppressants. These authors further state that withdrawal, regardless of the initiator, can lead to severe deleterious consequences for the health of the AIH patient. Because patient-initiated withdrawals are often not documented in the clinic due to the patient masking their noncompliance, the phenomenon is important to study as a contextual factor. Studies have explored the frequencies of patient-initiated withdrawal in various withdrawal studies, although the estimates vary highly. The most extreme documented rate of patient-initiated vs. all-cause withdrawals was 713/844 (84.5%) (Van Gerwen et al, 2013).[64] Hoeroldt et al (2011)[65], in a United Kingdom-based AIH study, noted that 29/84 (34.5%) of withdrawals from an AIH treatment trial were patient-initiated and due to side effects. The lowest rate of voluntary withdrawal was observed was 1/14 withdrawals (7.1%) (De Luca-Johnson et al, 2016).[66]

The physical environment, defined as the physical surroundings (including geographic surroundings) of the patient, is another often-studied contextual factor.[67] A meta-analysis by Wen et al (2018)[68] discovered that two geographic factors played a role in the mortality of hospitalized patients with AIH; patients residing in metropolitan areas of under 250,000 population experienced significantly higher mortality rates than those residing in large metropolitan

(population > 1,000,000) regions, and patients residing in the western region of the United States had significantly higher mortality rates than the nation overall. Substance exposure, another element of the environment, may also modulate the pathogenesis of AIH; Ngu et al (2013) [16] suggested a role for exposure to wood smoke (via wood-powered heating systems) in *decreasing* the future relative risk of developing AIH.

Environmental research in AIH has taken the first steps in extending beyond the physical environment and has also encompassed the psychological environment, in particular patient quality of life. A Russian study (Golovanova, 2010[69]) compared quality of life in AIH-PBC overlap patients across different medical treatments (combination ursodiol and prednisone vs. monotherapy ursodiol) and utilized the 36 item Short Form general (mental and physical; SF-36) survey to differentiate quality of life in these cohorts. This research concluded that combination ursodiol-prednisone therapy was associated with higher SF-36 scores (and therefore a greater estimated quality of life) than what was seen with ursodiol monotherapy. In addition, Srivastava and Boyer (2010)[70] conducted survey-based and qualitative psychology analyses on AIH patients and controls. Using the psychological stress instrument Social Readjustment Rating Scale (SRRS), they determined that while AIH patients did not score significantly higher than matched controls for psychological stress. However, when AIH patients with at least one relapse were compared to controls, it was noted these relapsing AIH patients have significantly higher psychological stress scores via SRRS. The qualitative analysis in this study revealed that stress experienced by AIH patients was due to reasons that were similar to that seen in the general population. The major difference in reasons for stress focused on daily living, with AIH patients expressing trouble with work, school, and exercise due to the fatigue inherent in the disease.

Exposure to other substances, including recreational drug products, may also modulate AIH's pathogenesis. Smyk et al (2012)[71] suggested that first-hand tobacco smoke exposure correlated with an increase in the development of primary biliary cirrhosis (PBC) and by indirect conjecture hypothesized that it may also play a role in the pathophysiology of AIH. A New

Zealand-based survey by Ngu et al (2013)[16], in contrast, rather interestingly discovered a decrease in the risk of developing AIH risk in those with a history of regular ethanol (beverage alcohol) consumption.

Finally, some articles subject to full-text filtering did not explicitly research contextual factors but advocated the use of contextual factors to supplement their research. Clinical research focused on *de novo* (transplant-induced) AIH (Visseren & Darwish-Murad, 2017)[72] has advocated for the addition of non-clinical factors into determining outcome influencers. Informatics-based researchers have also called on the addition of contextual factors to analysis: Sonnenberg and Naugler (2010)[73] called for a very detailed analysis of patient "social factors" (p. 722)[73] to add to their existing mathematical model of liver disease prediction, which only used clinical factors. Mells et al (2013)[74], after conducting a genome-wide association study (GWAS) to attempt discovery of AIH genomic correlates, found only weak genetic evidence and called for the incorporation of contextual factors as additional variables into GWAS and other bioinformatics-based genomic studies.

In addition to having potential contextual etiological factors, AIH, like other diseases, affects the whole life of the patient, including in non-medical (contextual) areas. Therefore, a search is undertaken to determine the impact of *quality of life* (QOL), a very important contextual factor domain encompassing psychological and functional facets of life, in AIH patients. The query *AIH quality of life* was utilized in PubMed and retrieved 19 results.

It was noted that the impact of AIH on global quality of life scales (those that combined physical, mental, and social aspects of QOL) was significant. Janik et al (2018)[75] noted that Short Form 36 (SF-36) global QOL responses were inferior in AIH patients versus controls. Via analysis of Modified Fatigue Impact Scale (MFIS) results, the impact of fatigue was seen to be higher in AIH than in matched controls. Psychosocial QOL factors were overall the least impacted. Furthermore, prevalence of anxiety and depression was found to be higher in AIH patients than in matched controls in this survey-based study. Objective signs, symptoms, and

comorbidities, including cirrhosis, pain, and jaundice, were also demonstrated to adversely affect QOL in a pediatric study; furthermore, this study found that severe grade AIH more negatively affected pediatric QOL.[76]

Psychologically-specific sequelae were noted in one study[75], which used the Personal Health Questionnaire 9th Edition (PHQ-9) to assess psychological QOL in AIH patients and also clinical correlates of QOL. General psychological QOL was adversely affected by female sex and usage of the drug prednisolone. A more specific and qualitative observation, furthermore, was an increase in anxiety regarding the stigmatization of AIH due to its hepatic nature and untoward associations with infectious and alcoholic hepatitides.

Overall, it can be concluded from this brief synthesis that contextual factors are a nascent, albeit insufficiently studied, modifier of AIH's patient course. Further research is therefore necessary to record and classify contextual factors experienced by AIH patients.

2.4 Patient Activation & Engagement: Social Media

In order to extend the case of AIH into that of social media research, I now turn to molding literature regarding an existing framework to social media. It is a generally accepted principle that patients are best to be involved in their own care (i.e., self care). This onus is most likely stronger in patients with rare disease, who in many cases have insufficient time to discuss their entire life situations with providers during brief and sparse office visits. For purposes of this discussion, patient activation and engagement (PA and PE respectively; here together PA&E) will be defined as a motivation for patient attitude in self care and the ability to change patient behavior, respectively. The two are interdependent and therefore treated as one concept (PA&E) here. Most importantly, PA&E will be explored as it relates to social media in healthcare in order to create a grounded initial case for studying social media for healthcare research purposes.

The terms as explicitly used were searched on PubMed, with a restriction to articles published in the past ten years. The query *"social media" AND ("patient engagement" OR "patient activation")* was utilized. Nine articles resulted, of which five were removed due to not discussing patient impact or due to being large-scale reviews. In this brief thematic synthesis, the four remaining articles are assessed to examine the effects of social media on PA&E.

Collier's article (2014)[77] took the form of an editorial covering the marketing aspects of pharmaceutical companies using social media to inform patients of products and help them manage disease. Liddy (2017)[78] covered an example of a Facebook page used to raise electronic primary care virtual consult (eConsult) awareness. Meanwhile, Rozenblum et al (2017)[79] discuss the current status of social media use in healthcare. Finally and perhaps most poignantly, Dhar et al (2018)[80] put forth an exploratory research study in the field of liver transplant (a procedure coincidentally performed on a significant proportion of AIH patients), where a controlled, closed Facebook group is utilized to facilitate patient information sharing.

The first theme noted in these articles is the *benefactor*. For purposes of this discussion, the benefactor is the primary party (outside the researchers and authors) benefitting from the social media venue(s) in question; ideally, one benefactor will be the patient. Significantly, Collier (2014)[77] noted that pharmaceutical corporations would be benefactors in social media usage, with (in this use case) patients taking a second priority. This editorial specifically noted that one scenario in which patients can be directly engaged (and therefore benefitted) is the provision of disease management portals (albeit ones that focus on the company's drug products). This observation is significant because in patient-centered medicine, patients are inherently the chief benefactors. Fortunately, in contrast, Liddy (2017)[78]'s study implies that patients will benefit from the eConsult services advertised on their Facebook page and could also leave feedback about the services. Dhar et al (2018)[80] were at the other pole of patient-centered approaches, allowing most discussion to be held between patients for purposes of mutual advice

and social support, leaving patients as virtually exclusive benefactors of the study; this structure is what is eventually observed in this dissertation's research.

The next theme to be explored is the *nature of information to be gained* by the benefactor (which in turn is defined as the previous concept). The nature of information to be gained is tied to the nature of the benefactor. Business-oriented organizations such as pharmaceutical corporations and eConsult services primarily gain information that is best categorized as feedback. Collier[77] notes that pharmaceutical companies stand to gain information about adverse drug events (ADEs; including side effects) and drug efficacy. Liddy[78], although implying more benefit to the patient, also notes that feedback on eConsult services is gained via consumer (patient) reviews, which can be posted to their social media venue. Dhar et al[80] and Rosenblum et al[79] both propose that patient contextual factors (social and environmental health-impacting factors not ascertained in the clinic) can be obtained from. Patient contextual factors, which surround the patient more often than the clinic does, intuitively will impact PA&E and clinician knowledge of these factors will be important in improving PA&E. Dhar et al[80] further find that (similar to Collier's[77] pharmaceuticals editorial) ADE and treatment efficacy opinions can be gained, and further notes that symptoms (related and unrelated to liver transplant) can be surveyed via their social media venue. However, Dhar et al[80] strongly imply that the information gained is to be used directly in the body of knowledge of engaging and activating patients, as opposed to simply a business strategy.

The final theme noted is the *dyad* of communication. The dyad is an ontological concept defined here as the subjects and objects of the verb *communication*. Greater patient involvement in communication dyads is a cornerstone of increasing PA&E.[81] Rozenblum et al[79] propose that the dyad is chiefly patient-to-provider (or patient-to-researcher) communication, forming unidirectional structures with the patient as subject. Collier[77] proposes that a communication dyad is created between patients and pharmaceutical companies. In the given scenarios, the dyad is bidirectional, with one party as subject and the other as object. Pharmaceutical companies are

expected to advertise product information to patients; on the other hand, patients communicate their opinions about the product to the companies responsible for the social media venues, so both actors in the dyad may be subject or object. Liddy[78] notes a similar bidirectional dyad, with organizations increasing patient awareness of eConsult services, and patients providing feedback about such services. PA&E here occurs most significantly in the form of increasing patient awareness about a service that can improve their health. In contrast, Dhar et al's[80] study notes that the primary observed dyad was patient-to-patient, with patients engaging each other. Occasionally, a patient education dyad of provider-to-patient would be observed, and finally, the dyad of patient-to-provider via research observation may increase the body of knowledge surrounding PA&E.

In this thematic synthesis, multiple themes affecting patient activation and engagement (PA&E) were noted across four relevant publications. The identity of the social media venue's *benefactor*, the *nature of information to be gained*, and the nature of the *dyad* of communication all affect PA&E, whether directly or indirectly. These and similar themes are important in assessing the role of social media in rare disease research; I will therefore address such items in the forthcoming sections.

2.5 Social Media: An Introduction

This dissertation earlier has noted that patient activation and engagement (PA&E) are amenable to social media use*;* here I provide a brief background prior to proceeding in deeper discussion on social media in rare disease research. Online social media, often known simply as *social media* or *SM*, encompasses a wide variety of Internet-based venues dedicated to the mutual sharing of information between friends and across the world. Many types of SM qualify and are more often known as *social networking sites* (SNSs) due to their ability to facilitate direct interactions between different users.

Facebook™ (http://www.facebook.com/)[82] is an SM venue used by over half of the US adult population and reaches penetrations of 73% among Internet-using adults ages 30-49. [83] As such, it is arguably the world's most popular social media venue. Facebook™ provides support for individual personal *walls*, where users can post things that are on their minds and also solicit reactions from individuals they have designated as friends. Posts on walls can also be shared globally if an individual chooses to do so. Facebook™ functionality is extended by *Facebook Groups*, where groups of people with similar interests can create a common wall on which to post and share thoughts.

In contrast, the online *blog* (short for *weblog*) is a service that is typically owned by private individuals. Most blogs focus on the owner's thoughts and writings, and also offer a place for viewers to react by *commenting*. The corporate run service Twitter™ is a form of blog (*microblog,* as posts are limited to 140 character *tweets*) that is popular for sharing brief thoughts on a global basis. [84] The final common form of online social media consists of *Internet forums*, also known as *web forums, online forums, message boards,* or simply *forums*. These venues are almost always privately run and allow for textually rich communication that is organized into posts (initial communications) and comments (reactions to the initial communications).[85]

2.6 Social Media for Studying Patients with Rare and Common Diseases

Initially, I conducted a brief systematic review of the literature to assess the current (2007-2017 as of review) state of social media in general health and biomedical research. PubMed was queried for the term *social media* (all fields; no quotes). Only results within the past ten years of query (i.e., those published during or after 2006) were included. Further filtering was performed to only include articles of types *study, comparative study, observational study,* and *clinical trials*, thereby excluding articles such as reviews and editorials. 568 articles were returned in the search results, which were downloaded in the form of PubMed/Medline XML.

XSLT (XML Stylesheet Translation) was used to parse out pairs of PubMed IDs and the titles that the IDs represented. I then performed a systematic qualitative analysis on the resulting titles, marking whether the title of each article indicated that it was actually an original study in the healthcare field that related to patients participating in online social media.

In order to determine relevant article titles at this stage, articles were excluded if it was clearly evidenced by the title that one of the following exclusion criteria were relevant:

- Criterion 1 (NHC; N=14 excluded): The title evidences a definite lack of relationship to health or healthcare
- Criterion 2 (NSTD; N=7 excluded): From the title, it is clear that the article is not an actual study (i.e., it is a review, proposal, or editorial)
- Criterion 3 (CSTD; N=52 excluded): The title clearly shows that the article involved only the study of clinicians (and not patients/lay consumers).
- Criterion 4 (NICT; N=364 excluded): It is obvious from the title that the article does not deal with information and communicating technologies (ICTs; virtually all of these articles dealt with non-ICT based materials including print-based, television-based, and face to face interventions).
- Criterion 5 (OWC; N=52 excluded): The title clearly shows that the study did not include patient-generated communication as valuable input; most of these were one-way educational platforms for communication of information to patients).

Of note, the criteria were sought in consecutive order during the review of each title, so if the article was clearly not about health (Criterion 1), the assessment for further criteria was not conducted. Hence, all of the exclusion criteria counts add up to the total number of titles that were excluded. The final title analysis therefore excluded 489 articles (82.6% of the articles returned by PubMed) and similarly included 79 articles (17.4% of articles returned).

I then performed further exclusion analysis via qualitative interpretation of the abstracts of the 79 remaining articles. The same criteria in the title analysis were applied in an identical consecutive fashion. One article excluded due to lack of relationship to health, 4 articles were excluded due to not being actual studies (NSTD), 2 articles were excluded due to being solely about clinicians (CSTD), 3 articles were excluded due to not having an ICT-related component (NICT), and 29 articles were excluded due to having no patient-generated communication component (OWC). Therefore, in the abstract analysis phase, 39 articles (49.4% of those that passed the title analysis) were excluded, while 40 (50.6%) of these articles were included.

Because the articles have been filtered to only the ones that will be relevant for future discussion, a narrative review can now be performed. The knowledge gap is evidenced in that the systematic review demonstrated that there were only 40 articles published between 2006 and 2016 that fit the criteria of engaging and researching patients via social media tactics. The next and final step in filtering involved judging by the article introduction full text sections if the article focused on rare diseases. Out of the 40 articles, eight did, an amount sufficient for a narrative review, which is discussed further.

Given the wide reach and content diversity of social media, it is not surprising to note that several studies have shown that social media is popular in the rare disease community in attracting and coalescing groups of rare disease patients. Significantly, the social media venue Facebook™ has shown a very strong trend towards attracting those with rare diseases, [7, 86] and part of this trend is, as hypothesized by Davies (2016)[18] in one proposal on SM use in Rare diseases, due to the fact that there are 1.7 billion individual (non-group) user accounts on that SM venue.[87]

More specifically, Schumacher et al[86] detail the problems of finding patients with post-Fontan procedure plastic bronchitis (PFP-PB) and protein-losing enteropathy (PFP-PLE), two rare complications of a surgery (the Fontan procedure) that itself is rarely performed. These authors were able to discover via Facebook™ and online surveys what is believed at the time to

be the world's largest ever study cohort of PFP-PB and PFP-PLE patients. Similarly, Greeley et al (2011) discovered, researched, and have been following up with a cohort of over 700 patients recruited from various Facebook™ patient support groups; these patients had monogenetic diabetes, a rare subtype of the disorder.[88]

Health-related information is commonly shared across rare disease-related online SM venues; therefore, determining the types of health-related information shared is important to not just research AIH, but also to determine the types and quality of information that are shared on the AIH-related SM venue of study.

Rare disease-related blogs (a subcategory of SM venues) often focus on material that directly relates to health information. Evidence exists in using SM as a venue for clinicians to disseminate useful, reliable health-related information to patients, lowering barriers to clinician-patient communication. Hawn[89] has proposed that patients are directly engaged, taking the first step to be activated in their own healthcare, by using social media as a venue of receiving health-related information. One study found that blogs run by parents of children with Hirschsprung's Disease (HD) were effective in answering questions of fellow parent caregivers and moreover, in quickly attracting thousands of views from around the world.[90] One physician has found that the blog is a highly effective medium over which to distribute information and answer patient questions about pheochromocytoma, a rare endocrine tumor.[91] In another observation, a group of PF-related blogs allowed for high volumes of patient communication, and that the most popular topic in patient-generated content related to the symptoms patients were experiencing.[92]

The microblog service Twitter has also shown to be of use for increasing societal awareness of the rare blood cancer blastic plasmacytoid dendritic cell neoplasm (BPDCN), as shown by Pemmaraju et al (2016). This study has shown that multiple types of health information can be disseminated over a Twitter-based system; recipients of information are patients, clinicians, and other stakeholders concerned about the disorder.[93]

An overall survey of social media sites performed by Winchester et al[94] sought to describe the quality of information regarding cardiac stress testing disseminated over these venues. These authors found that a commonly discussed topic was the *Choosing Wisely* campaign, a clinically-approved movement that advocated for patients and doctors to wisely choose whether or not stress testing was necessary. This study therefore demonstrated that information regarding proper healthcare protocols was being disseminated to patients.

Finally, it was also found that an online focus group geared towards patients with rare cancers allowed patients to discuss their issues in a more honest fashion. One specific benefit of using an SM venue was noticed as the fact that patients, as users, were *masked* by screen names and did not reveal their true identities, allowing for greater transparency. [95]

2.7 Online Patient Support in Rare Disease

Patients with rare diseases, including AIH, travel great distances to seek presence of a provider, both for treatment and advice. [4, 5] It was noted in the previous chapter that contextual factors also related to the ways in which patients emotionally behave and react to their disease. Knowing these facts, it is understood that patients with AIH (and other rare diseases) will inherently face a shortage of support, both socially and informationally, and literature indicates that patients of many rare diseases turn in addition to their peers for support.

Much prior research on peer support over social media has been performed over online health communities (OHCs), which for this purpose are defined as Internet-based communities of individuals affiliated as patients or caregivers of various disorders. It is theorized that the exchange of support of all types is facilitated by the *post-and-response* (or in Facebook's case, *post-and-comment*) nature of the OHC.[96] Significantly, the social media (SM) venue Facebook has shown a very strong trend towards attracting those with rare diseases, [7, 86] and part of this

trend is, as hypothesized by Davies (2016)[18] in one proposal on SM use in rare diseases, due to the fact that there are 1.7 billion individual (non-group) user accounts on that SM venue.[87]

From an analysis of the literature, two types of (emotional/social and advice/informational) as well as two directions of (inbound and outbound) were noted. Wang et al, furthermore, proposed a comprehensive, computationally-based classification system of the types of social support given over OHCs; these authors propose that seeking and providing informational support, seeking and providing emotional support, and seeking and providing companionship (ie, social/emotional support) are major areas in social support offered over OHCs.[97]

Advice and/or informational support (here abbreviated AIS) is defined here as the provision or receiving of objective information or suggestions as a form of help. The presence of this type of support is particularly important because rare disease patients have relatively sparse visits with their providers and require other sources of information. PatientsLikeMe, a forum-based website for patient support, was observed to be useful in facilitating exchange of knowledge about treatments. [98] In a qualitative review of pulmonary fibrosis (PF)-related blog postings, sharing health-related information as support was reported to be the most common type of support found. [92] One study found that blogs run by parents of children with Hirschsprung's Disease (HD) were effective in answering questions of fellow parent caregivers, in the process providing informational support.[90] In addition, it has been found that patients with rosacea, a relatively uncommon skin disorder, actively seek information and advice through online health communities.[99] More saliently, a publication on online health communities catering to rare vascular diseases claimed that AIS was the most important kind of support exchanged over rare disease-oriented online health communities due to the inherent lack of availability of external advice and information regarding such conditions.[100]

Social/emotional support (here abbreviated EMO) is considered here as support that does not have objective information or recommendations but instead are essentially kind words given

to help someone's emotional status. Inherent in rare and incurable diseases is grief that the body will never return to normal; indeed, a study regarding online support in grief discovered support content to be almost entirely emotional in nature.[101] Similar findings were obtained in a study of another psychological problem, postpartum depression.[102] PatientsLikeMe in one case was also found to allow for the exchange of emotional support for patients with various uncommon (and common) chronic conditions. [98] The aforementioned study of PF-related blogs found that the second most popular type of discussion (behind sharing health-related information) involved the solicitation and provision of social (ie, emotional) support. [92] A study from Spain, furthermore, found that Facebook pages that are administered by official rare disease advocacy organizations also facilitate the provision of information as well as social support to patients. [7] A Dutch survey on breast cancer patients and survivors seeking online support, meanwhile, reported that emotional and social support predominated among the support needs and capabilities of affected women.[103]

Support can also be designated not just by subject type, but also by its direction and structure, although this aspect of support is less well-studied. In another breast cancer study, the provision (here known as *offering* and abbreviated OFF) and acceptance (the initial stage, *requesting,* abbreviated here as REQ) of support were demonstrated, although the utility on psychiatric clinical outcome of the patients who provided support was debatable.[104] A Chinese study of online Acquired Immunodeficiency Syndrome (AIDS) patients on the microblogging service Weibo demonstrated the importance of dyads in provision, acceptance, and reciprocation of all types of disease-related support.[105] A contingency management intervention in tobacco smoking cessation demonstrated the importance of bilateral support in disease management, especially in situations such as contingency management here group reward and punishment are present.[106] Finally, a computational linguistics-based study of a Norwegian online support group demonstrated that the group's content changed semantically over time to represent a more egalitarian structure of support, which was considered a positive development by the authors.[107]

2.8 Social Media & Autoimmune Hepatitis

Autoimmune hepatitis (AIH), being among the rarest of rare diseases, inherits from other rare diseases all onuses for social media-based study; indeed, this condition has a significant online support community. A colleague of the author's, who cannot be named due to concerns over his patients' privacy and will be referred to as "AC" (*administering colleague*), is a hepatologist who administers an (also unnamed for the same reason) Facebook™-based group that allows informational communications between researchers, patients, and clinicians. During the time span between May 2015 and May 2017, 1,052 users (excluding the administering colleague) have been active in this group.

In his experience moderating this group over the past three years, the administering colleague believes that the majority of users of the group are actual AIH patients or caregivers for those with AIH. The primary author, furthermore, has performed preliminary computational topic modelling studies on the content generated by the group's users over the two-year timespan. These studies have shown that topics represent discussion of personal histories of disease, treatments, and also the provision of social support. Most importantly, this group comprises corpus of users and communications that will be of study in this dissertation.

2.9 Barriers to Health Research via Social Media and Overcoming Them

In light of the desire to study AIH via social media-based methods, caution must be exercised to potential barriers to researching health via social media; these may broadly be classified into technical and social/ethical barriers. Intuitive concerns exist because the proposed

research involves acquiring large quantities of unstructured, personally-identifiable user-generated content from a website with proprietary access controls and structuring it in a manner that could allow those with access to the resulting structured data to easily identify individual users.

Technical barriers exist, and these barriers are most often socially-grounded. Since 2010, Facebook™ has banned a majority of automated information collection (*scraping*) via application program interfaces (APIs), which used to be a mainstay for mining Facebook data.[108] Extant Facebook Apps (applications) and application programming interfaces (APIs), including the Graph API[109] and content feeds allow for some collection of user information, but not universal access to consenting users' private, full text communications. Privacy and ethics issues also exist: Regardless of how the data are collected, Facebook™ screen names are valid e-mail addresses and considered identifiable private information (IPI) by most institutions, including that of the authors'.[110] In addition, users' expectations of privacy on Facebook™ Groups may vary, especially if the group content is set to *closed* (i.e., only admitted members may view and create group content). Given that back-end downloading is prohibited, researchers may only work with the source code of the website (that which is available from any common web browser). This fact implies that researchers must work to determine the regular source code expressions in which the necessary data and metadata are stored so that users can be related to their communications and to each other.

A review by Abedin et al (2017) [111] surveyed literature regarding social media use for research; this review synthesized that researchers and SM users believe that de-identification of information is mandatory in order for such research to be ethically and legally conducted. A general study of the ethics of SM information extraction, furthermore, found that the use of informed consent is important, particularly in utilizing information that users may believe has a modicum of privacy.[112] (For the reader's reference, a Facebook example of such information

would include that present on a *private group*, which is a group that is only accessible to those whom an administrator approve.)

However, even users on publicly accessible social media venues have experienced objections to being studied. A seminal 2004 study by Hudson & Bruckman[113] discovered the pitfalls of using publicly accessible but moderated *chat rooms* (SM venues where users could communicate in real time by typing text to each other): After entering a chat room, the researchers would announce their identity as researchers; in over 63% of chat rooms studied, the researcher was subsequently ejected and *permabanned* (permanently banned) from the chat room by a moderator. Nonetheless, with disclosure of researcher status and absent any external moderator or administrator objection, both of which are precautions being taken in the course of this dissertation's research, it is noted that research of the SM venue is much more acceptable to users. [111]

Furthermore, specific local policy to the author is mediated by the organization (Indiana University Institutional Review Board; IU IRB) responsible for approving the protocols in this study. Specifically, the IU IRB classifies Facebook™ screen names as identifiable private information (IPI)[110], that which is not protected health information (PHI) but whose release would allow subject identification outside the social media venue in question. In particular, the IPI classification is given because these screen names are actually valid email addresses that link to the users' respective Facebook messaging accounts.

Without overcoming the barriers explained just prior, utilizing social media data to elucidate AIH would be unethical or entirely impossible. There are, however, ways to mitigate these barriers with appropriate technical and ethical controls. Therefore, I developed and executed a front-end data mining method to acquire data from a colleague-administered Facebook™ group pertaining to AIH, and further describe security and de-identification protocols to remain within bounds of commonly accepted research ethics. Analysis of the extracted data

then confirms many of the putative benefits of using social media in order to research "digital cohorts" [85] (p. 614) of rare disease patients.[114]

Finally, the group comprises an excellent convenience sample that may in fact encompass a significant portion of the global AIH population. Being led by researchers, these groups are open to observation for research without any risk of the researcher being ejected. The chosen group, with the moderators (including AC) taking active roles in preventing *spam* posting. Users in these groups communicate openly knowing that only fellow group members and researcher administrators have access to the content they post in context of their personal identities.

Note: Specific technical data mining techniques and ethical measures, including consent protocols and IU IRB approval, are entertained in forthcoming Chapter 3 of this dissertation.

CHAPTER 3. ACQUISITION, STRUCTURING, AND PROTECTION OF AIH-RELATED FACEBOOK DATA

3.1 Introduction

As discussed prior, utilizing social media data to elucidate AIH would be unethical or impossible without adequate technical methodology and ethical controls. Technical issues, many of which are socially-grounded, exist; furthermore, there are ethical issues that relate to the usage of identifiable private information (IPI).[85]

Therefore, I developed and executed a front-end data mining method to acquire data from an IU Health-sponsored Facebook™ group pertaining to AIH, and further describe security and de-identification protocols to remain within bounds of commonly accepted research ethics. Analysis of the extracted data then confirms many of the putative benefits of using social media in order to research "digital cohorts" [85] (p. 614) of rare disease patients.

3.2 Methodology

Facebook codes its front end (web interface) in XHTML (eXtensible HyperText Markup Language). Because back-end downloading of such detailed Facebook data is prohibited, data were acquired by the entire group page into the browser front-end. Posts on the page were then expanded to reveal all comments by using an open source JavaScript bookmarklet. Finally, the browser's *view source* feature was used to download the full XHTML source code. The XHTML file's document object model (DOM) was elucidated manually to determine the locations of posts, post authors, post timestamps, post comments, post comment authors, and post comment timestamps. All were found to be contained in regular XHTML expressions, allowing for canonization of the form.

A novel canonical metalanguage, the Facebook Metalayer Language (FML), was therefore created for this dissertation's research in order to exclude visual element coding and compactly store only the data that are needed for text and network analysis. The XHTML-to-FML translation was performed using XML Stylesheet Literal Translation (XSLT) via Oxygen XML software.[115] Detail and flows are shown in Figures 1 and 2, respectively.

Specifically, FML contains only:

- Posts: ID, content, author, timestamp
- For each post, comments: Content, author, timestamp

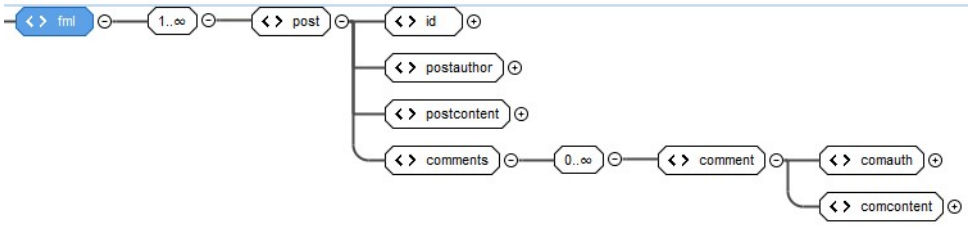

Figure 1. RELAX NG (RNG) Metalanguage Schema for FML.

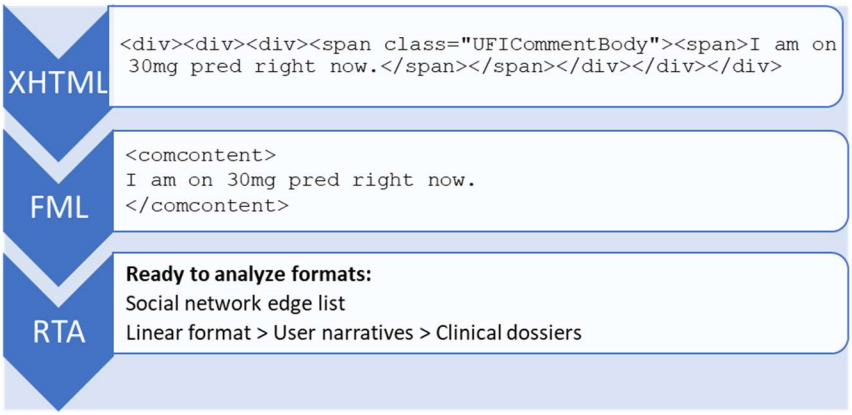

Figure 2. The XHTML to FML to Ready-to-Analyze Format Workflow

For future natural language processing (NLP) analysis, the FML file was parsed via a third XSLT sheet into a tag-stripped delimited linear format, with one line for each communication and its metadata. In order to compare communications throughout the population as they differed between separate users, the messages were then into files, with each file representing every communication a single user made over the two-year course of observation. These files are termed *user narratives,* parsed from the linear file; one is visible in Figure 3. Timestamps were also parsed into the number of days past 2014 January 1.

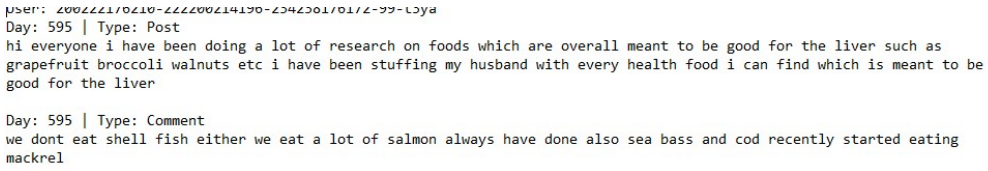

Figure 3. Partial User Narrative Screenshot

Ethical concerns are equal to technical ones. The group and its administering clinician have already gained informed consent to research user communications via a front-panel agreement stating that users have no expectation of privacy when posting in the group and that communications may be used for research purposes. However, to avoid further issues with privacy and ethics, the following internal study policies were devised by the authors and approved by the Indiana University IRB:

- Any data containing identifiable private information (IPI), including Facebook screen names (which are valid email addresses and cognate to real life names), are always stored on encrypted media.
- Any data that leaves an encrypted drive are always de-identified via user name encoding to remove PII.

- For purposes of research integrity and to comply with Facebook's terms of service, the researching author (AK) is never allowed communication with members of the group.
- The clinician co-author (AC) is not allowed to view the key that links the Facebook screen names of users with their encoded names.

For de-identification, all usernames in the FML file were encrypted numerically by using a Java Virtual Machine (JVM) program and fixed key, kept on an encrypted hard drive. A fixed key was chosen in the event that more data from the group were to be added, thus permitting referential integrity between group users' posts.

3.3 Results

Front-end downloading was utilized to acquire 73.1 MB of XHTML data, representing all group communications between 1 June 2015 and 31 May 2017. Via removal of unnecessary (visual) XHTML elements over XSLT, the XHTML was condensed into 6.6 MB of FML.

The integrity of the XHTML-to-FML conversion was validated by randomly sampling 100 user communications from the FML. Communications were then searched via the encoding key through the XHTML to find whether the user who generated the communication matched between the XHTML and FML. 100 (100.0%) of all communications matched the user across both formats. The linear file was validated similarly; this time, the linear file and FML file were compared, with a match rate of 100.0% as well.

3.4 Discussion & Conclusions

A proof-of-concept in acquiring user-generated content from an active Facebook™ support group for a rare disease, autoimmune hepatitis (AIH) has therefore been demonstrated.

Importantly, the required data gathered via a method (front-end download) that, as of this writing, is not prohibited by Facebook's terms of service. Downloaded data were successfully parsed into a more compact form (Facebook Metalayer Language; FML) that maintained full information integrity of communication content via metadata. The resulting data are therefore usable in studies of online posting patients with AIH, and these protocols may also be used to gather patient-generated online content from users with other disorders of interest.

CHAPTER 4. DETERMINING HEALTH INFORMATION SHARING OVER AIH-RELATED SOCIAL MEDIA

4.1 Introduction

The data acquired in the previous chapter will be subject to meaningful analysis in order to attempt detailing the clinical and non-clinical histories of AIH patients. Health-related information is commonly shared across rare disease-related online SM venues; therefore, determining the types of health-related information shared is important to not just research AIH, but also to determine the types and quality of information that are shared on the AIH-related SM venue of study. [85, 116] A model thereof is conceptualized in Figure 4.

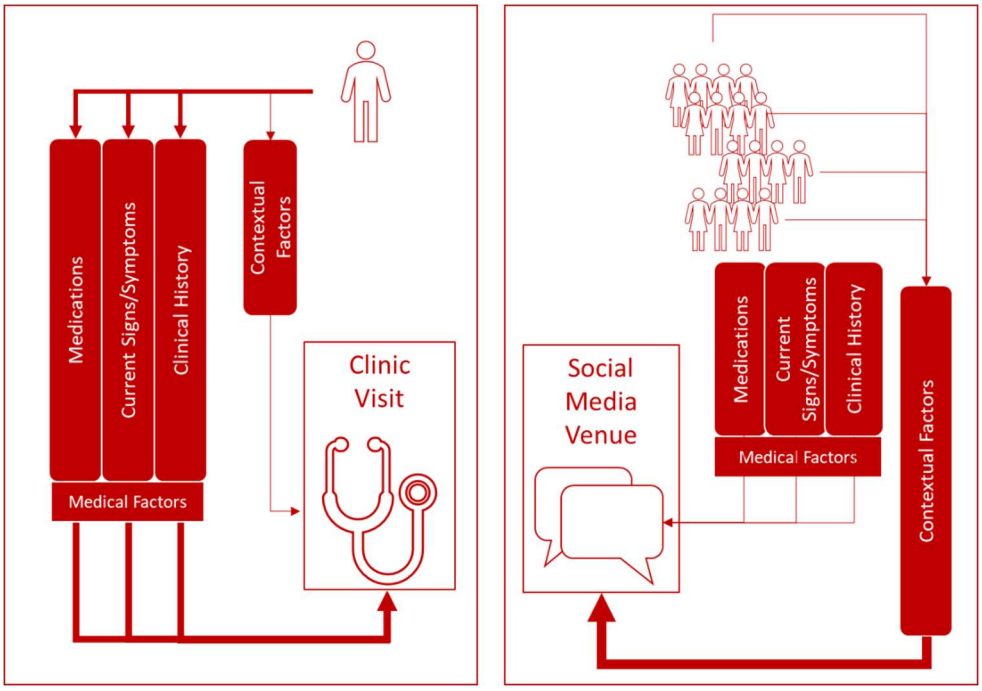

Figure 4. Conceptual Graphic of Health Information Shared: Clinic vs. Social Media

4.2 Methods

The data extracted and used in earlier research[117] were re-used for purposes of determining health information sharing over the venue. Importantly, I utilized the *user narratives*, the latter of which each contained the entire communications of its respective user. In order to quantify communication structures within this Facebook group, social network analysis (SNA) was performed by transforming the FML into a network (graph) edge list. SNA itself was conducted using Cytoscape.JS[118] software. Edge lists consisted of a giver (commenter) and receiver (author of post being replied to). To confirm reliability of actual user communication (as opposed to the reliability of computational methods for mining and transformation), the user narratives of 73 randomly selected users were qualitatively annotated to estimate the role of the user (patient, caregiver, or other) and gender of user (or patient user represented). The suspected residency of the user (US, non-US, or ambiguous) was also noted by the usage of US brand names for medications as well as US word spellings.

Content was also assessed using a deeper qualitative analysis. Six (6) randomly selected user narratives from the previous pool of 73 were annotated for contextual factors and organized with the framework proposed by Holden et al.[67] These narratives were then detailed further into *clinical dossiers*; items such as clinical information, demographics, and family history were noted; finally, contextual information gained in the step prior was added to the dossiers.

In order to give an all-encompassing yet quantitative view of the topics discussed, and therefore types of information shared, over the group in question, topic modelling (TM) via latent Dirichlet analysis (LDA) was performed. TM via LDA uses an *N-incomplete* algorithm to efficiently sort *baskets of words* that define the similarities and differences between documents in a given corpus. Documents are then clustered and *key topics* representing the word baskets are assigned a relative strength to each document, creating a document-key topic strength matrix.

The LDA implementation used in this case is the Machine Language Learning Toolkit

(MALLET).[119] Most recently, Jones et al (2018)[120] noted that MALLET-based TM was of great utility in analyzing a similar corpus, that of breast cancer survivors over an online forum; these authors discovered both expected as well as novel topics of patient communication.

In this case, the corpus consisted of the previously generated *user narratives*, allowing to determine the strength of machine-generated topics across group users. After the cleaning of the corpus via stem word removal, MALLET was run over the corpus and instructed to return 50 key topics embodied by respective baskets of words. Qualitative professional interpretation of the topics was then performed by committee member AC, who is an AIH-treating hepatologist. If multiple key topics were found to imply discussion of similar matters, AC grouped these key topics into topic groups.

4.3 Results

Group metrics are then calculated via the resulting FML file; these numbers are shown in Table 2. Note that users who neither posted nor commented are excluded from this and further analyses because detecting them is physically impossible.

Table 2. Group Metrics

Active Users	Total	1053
	Who made at least one post	478 (45.4%)
	Who made at least one comment	1016 (96.6%)
Posts	Total	1479
	Made by AC	71 (4.8%)
Comments	Total	16646
	Made by AC	488 (2.9%)
	Received by AC	1798 (10.8%)

Qualitative annotation analysis of 73 randomly selected user narratives revealed the following user demographics, as per Table 3.

Table 3. Demographics from Qualitative Annotation

Role	Patient	Caregiver	Ambiguous**
	61 (83.6%)	10 (13.7%)	2 (2.7%)
Gender	Male Pt./CG of Male Pt.	Female Pt./CG of Female Pt.	Ambiguous**
	5 (6.8%)	66 (90.4%)	2 (2.7%)
Residency	Evidences US Residency	Evidence non-US Residency	Ambiguous
	55 (75.3%)	8 (11.0%)	10 (13.7%)

It was also noted that *contextual factors* (those unseen in the clinic) exist in the user narratives and add a vital layer of richness to the clinical details of patients. Six (6) randomly-selected narratives, referred to here alphabetically, were annotated for contextual factors and these factors noted in Table 4 below using the contextual factor categorization framework of Holden et al.[67]

Table 4. Contextual Factors Across Six User Narratives

Factor Category	User: Detail	Factor Category	User: Detail
Economic	C: Currently employed D: Employed in medical field	Psycho-logical	C: Considers self to be religious C: Can understand some scientific literature D: Considers self highly health literate D: Advocate for rare disease patients and causes
Health Behavioral	C: On a low-glycemic diet D: Using a low-carbohydrate diet D: Trying to lose weight F: On gluten-free diet F: Abstains from carbonated beverages F: Abstains from alcohol	Social	A: Has a daughter and granddaughter B: Has a young child B: Is married C: Has grandchildren E: Is married E: Has a daughter
Functional	C: Difficulty raising kids due to fatigue D: Recently able to take a camping trip	Techno-logical	C: Uses Internet to search for AIH-related study information F: Good at and enjoys using Facebook
Healthcare System	A: Specialist doctors of user are not communicating well and disagreeing with each other D: Has difficulty getting appointments	Other	F: Has a puppy

Overall, 24 contextual factors were discovered across the six users. These were also used to enrich the clinical dossiers; User F's clinical dossier is available in Figure 5.

Demographic	Age [redacted]	Gender: Female	Lives in [redacted]
Medications	Present • Prednisone • Mycophenolate • Probiotic NOS • Pantoprazole • Loperamide • Obeticholic Acid	Past • Ursodiol • DayQuil • Azithromycin • Triamcinolone • Fluticasone • Cortisone intra-articular injection	
Issues & Procedures	Present • Cirrhosis • Penicillin allergy • PBC/Overlap • Excipient allergy NOS • Droopy eyebrow • AIH stage: IV • Pruritus	Past • Pre-cancerous moles • Influenza • Injured ACL (knee) • Cholecystectomy	
Contextual Factors	• Abstains from carbonated beverages • Gluten-free diet • Has a puppy • Appears to be savvy using Facebook • Abstains from alcohol		
Family History	• Nothing noted		

Figure 5. Clinical Dossier Example

The rater (clinician co-author AC) determined that the clinical detail presented in these dossiers was less than what is typically gained during an initial clinic intake appointment but stated that sufficient information for research purposes was gathered; furthermore, he believed it a major benefit that analyzing group data would bring patients (users) that his clinic would never be able to see in real life.

Topic modelling with LDA through MALLET resulted in 50 key topics represented by respective baskets of words. Below in Table 5 is represented an example of one of the key topics' basket of words and the interpretation of its topic group by AC, and strengths of topic categories are noted in Table 6 thereafter.

Table 5. LDA-Generated Topic Example with Interpretation[116]

Key Topic Word Basket	AC Group Interpretation
euro acirc checked brvbar rosacea eye supplements set link gel hole pressure draw intolerant herbs metals janet hair broken bleeding	Alternative Treatments

Note: A list of all topics with key word baskets and AC's assigned topic category is available in Appendix I.

Table 6. Strengths of Topics and Categories Across the Corpus[116]

AC-Assigned Category of Topic	**Sum of Topic Strengths in Category**
Treatment side effects	23.23%
Caregivers	16.49%
Treatment stories	16.03%
Treatment goals	8.73%
Comorbid conditions	7.69%
New diagnosis	5.86%
Support groups	4.88%
Research	4.64%
Alternative treatment	4.28%
Religion	1.47%
Disease associated phenomenon	1.32%
Pathogenesis	1.24%
Disease side effects	1.11%
Pregnancy	0.74%
Treatment - pediatric	0.70%
Medication payment	0.64%
Research trials	0.21%

4.4 Discussion & Conclusions

Group metrics revealed the remarkable sample size gained from the data acquisition and analysis. During the initial qualitative analysis of 73 users, 55 users who live in the US were detected; this figure is over 1% of the total (less than 4,000) number of individuals in the US who

are diagnosed with AIH. If the fraction is extrapolated (to 1,052 users, matching 792 as US residents), nearly 20% of the US population with AIH will have been observed solely through the data that were obtained. A cohort of such size and coverage would be likely impossible to gather in the real world.

A general sense of the health-related information shared in the online support group at study was also ascertained. It appears that users are concerned with side effects of AIH treatment, because 23.23% of the topic weight centered upon this facet of discussion. The topics of communications generated by caregiver (as opposed to patient) users and general treatment stories each held approximately 16% of the weight of discussion.

Although the quantity and quality of medical factors found in the clinical dossiers, by AC's assessment, were weaker than what can be assessed in the clinic, the geographical reach is far greater and medical factors in this case can be ascertained for individuals who would never even be seen in the clinic. While it is valuable to gain confirmation of clinical insight from users, it is more noteworthy that contextual factors were elucidated. From the sample of clinical dossier users (N=6), 24 contextual factors were found. These factors, in particular those categorized as social and psychological, are very unlikely to be elicited in the clinical visit setting and provide richness to the histories of patients and the AIH population as a whole.

Inherent limitations of social media research exist. User communications are taken at face value and assumed to be honest and not fabricated. The limitation is inherent in social media research, although it should be observed that patients can fabricate personal histories in the clinic, as well. Furthermore, social media venues (including Facebook) prohibit page administrator contact with users (except through surveys where users can voluntarily opt in to contact), creating a challenge in connecting individual users' social media data with official clinical information. It is therefore recommended that health-related data from social media is better applied to population-level health rather than individual-level health research.

It is also noted that the demographics of the population at study may not reflect the demographics of the AIH population. The gender ratio (an estimate of the sex ratio) in the group's patients was estimated in related research[116] by the author to be 13.2:1 (female); most traditional studies suggest a sex ratio of around 3.6:1 (female).[121] Although 75.3% of annotated narratives evidenced US residency[116], AIH has shown a significant prevalence in other countries such as South Korea[122], Norway, and Sweden.[121] These limitations are expected to be inherent in any study involving online social media, whose content is considered dominated by females (with an estimated ratio of 1.5:1 for online support groups[123]) and English speakers.[124]

The dominance of the coauthor-administrator (AC) is discussed, particularly because he was the most prolific poster and his posts were subjects of 10.8% of comments received. The intense degree of gravitation could raise the speculation that discussion is driven into reflecting medical and clinical information rather than contextual factors; however, this conjecture is not sufficiently provable and it is further noted that that contextual factors were successfully ascertained in other cases.

The most important points of discussion pertain to future research: The purpose of this phase of dissertation work was only to determine proof of concept of extraction and estimate the reliability of user-generated content from this Facebook support group. The current research has been on a relatively small sample (73 or less) of 1,052 users. Future research will include analysis of clinical and lay terminology indicating contextual factors, medication, and signs and symptoms experienced, resulting in socio-medical *folksonomies* where user misspellings and slang usage are accounted for. With this *folksonomy*, automated term extraction via natural language processing (NLP) may be executed, allowing for large-scale mining of user/patient-generated online content for knowledge that will inform workflows of researchers and clinicians.

Overall, the research at hand shows great promise in benefitting the rare disease community's bodies of clinical and patient knowledge. Because these patients are difficult if not impossible to reach in quantity in real life but yet often post generously about their condition(s)

on social media, the ability to ethically extract and utilize social media information generated by cohorts of such users represents significant progression in rare disease research. Finally, it is noted via qualitative analyses that this user-generated content was demonstrated to provide clinically actionable information regarding the medical, and more importantly contextual, factors experienced by patients and populations with rare diseases including AIH.

CHAPTER 5. DETERMINING THE FEASBILITY OF DETECTING SELF-REPORTED XENOBIOTIC (DRUG) USAGE

5.1 Introduction & Background

Having obtained a suitable corpus of AIH patient-generated communications from social media in the previous chapters, an analysis of clinical factors is in order; I proceed first to discuss pharmaceutical medication drug products, also known as *xenobiotics*. The liver is a hub for processing xenobiotics and subject to damage caused by said chemicals, and AIH patients consume xenobiotics in the form of prescription therapeutics in order to control their disorder. Finally, in order to survive, humans (as with all living organisms) must consume some form of xenobiotic material in the form of vitamins and minerals.

I have already noted that the research body believes that various classes of antibiotics may mediate AIH risk. [16, 17] Statins, a commonly-prescribed class of drugs for high cholesterol, were also discussed as potential culprits. [16, 18-20] Other medications implicated by the literature in AIH pathogenesis include nitrofurantoin, carbamazepine, chlorpromazine, methotrexate, valproates, and even the common analgesic ibuprofen. [21, 22] Furthermore, the development of syndromes biologically similar to, and considered partially overlapping with AIH, are noted in some liver toxicity cases regardless of the alleged causative xenobiotic.[26, 27,28] AIH-associated biological liver damage features are also correlated with the use of non-prescription xenobiotics in the form of herbal supplements, including black cohosh[23], ginseng, ephedra, chamomile, milk thistle, *kava kava*, and pennyroyal. [22]

Xenobiotics in the form of pharmaceutical products *prescribed* for AIH treatment include a wide array of immunosuppressants, including corticosteroids.[30] The high side effect burden of immunosuppressant treatment may lead to an even higher load of xenobiotic burden when medications are added to control adverse drug effects (ADEs). [29]

The progress of researching AIH, including its pharmaceutical and xenobiotic demographics, has been severely hampered by an inability to recruit sufficient numbers of patients for studies; this inability is inherent due to the rareness of AIH. Preliminary content analysis via computational topic modeling has revealed that users in this group frequently discuss medications of intake as well as the side effects thereof.[116]

Therefore, I now explore the use of social media venues to study larger populations of AIH-affected individuals to determine the types of pharmaceutical products and dietary supplements that they claim to take.

5.2 Methods

Using nth-member sampling (n = 10), 105 user narratives (created as described in Chapter 4) were sampled via alphanumerical order of encrypted Facebook™ screen names. These narratives were then randomly ordered and split into 73 *training set* entries and 32 *testing set* documents.

The lead author (AK) and a coauthor (JSP), both with extensive academic natural science backgrounds, annotated the 73 training set narratives in order to observe mentions of medication-related terminology. For each post or comment made by each user, each volunteer was to note the following aspects:

1. *Literal.* The literal independent clause showing that (2), (3), and (4) were evidenced in the user's communication. This will include literal evidencing of xenobiotic use and therefore any misspellings that may occur.
2. *Xenobiotic.* Shows the actual xenobiotic drug. Components:
    a. Name of xenobiotic (US generic/USAN)
    b. Unified Medical Language System (RxNorm Concept Unique Identifier (RxCUI) system.[125]) corresponding to USAN

3. *Experiencer.* Categorizes the **mention** of a xenobiotic (not the user themselves) as to the user's use of the xenobiotic-related term. Classes:
    a. Patient
    b. Caregiver/relative of patient
    c. Observer: Mentioning xenobiotic but not due to taking it (e.g. as a suggestion or question).
        i. If the experiencer in the sentence is judged as observer, the annotation is omitted.
        ii. Non-observer users who act as observers will have only the observer annotations omitted.

Note: Users may have mentions of multiple distinct xenobiotics, and in many cases, the redundant mentions of the same xenobiotic.

4. *Temporal-Negation* (not recorded if Experiencer is an observer). Classes:
    a. Currently taking
    b. Taken in the past
    c. Denies taking (negation)
5. Type of spelling (annotated only by AK). Classes:
    a. Generic name, spelled correctly
    b. Generic name, clinical abbreviation
    c. Generic name, non-clinical abbreviation or misspelled
    d. Brand name, spelled correctly
    e. Brand name, abbreviated or misspelled
    f. Class of drug, spelled correctly
    g. Class of drug, misspelled

6. Literal split. From (1) above, split into – Components:
    a. Subject-verb portion
    b. Object (xenobiotic) portion

A consensus procedure was performed after annotation to ensure 100% agreement of the two annotators respect to each xenobiotic as mentioned by each user. The 35 testing set narratives were also annotated with post-annotation consensus, but without noting misspelling class. Therefore, a *folksonomy* of user-generated terms used to represent drugs and their intake can be generated.

The folksonomy, in turn, can then be structured in a set of *synsets;* the synset is a concept referring to the matching concepts with all possible synonymous (folksonomic) terms. [126] When a set of synsets (i.e., a thesaurus) is used to search a group of documents, the process is known as *named entity recognition (NER).* Synsets consist of one preferred term and multiple synonymous terms; the latter of which are misspellings, abbreviations, and synonyms of the preferred term. If a document (in this case, a user communication) mentions any member of the synset, the document (or communication) is tagged with having mentioned the concept represented by the preferred term.

The pharmaceutical compound *dexamethasone* (a xenobiotic used to treat AIH) can similarly be represented as a synset, with *dexamethasone* itself being the preferred term because the USAN nomenclature was used in earlier annotation, as seen in Figure 6.

Dexamethasone: _dexa_; _dex_; decadron; dexamethasone; intensol

Figure 6. Synset Representing the Pharmaceutical Dexamethasone

A xenobiotic folksonomy is therefore created via addition of drug name misspellings and brand nomenclatures are matched as synonyms to the respective generic compounds. [127, 128] A

combination of NER with rules-based searching further enhances the accuracy of detecting whether a xenobiotic drug is mentioned as actually taken by the user or just mentioned in passing (as an observer). Fuzzy matching via basic local alignment search (BLAS)[129] may have the capability to reveal misspelled xenobiotic generic and brand names. Furthermore, portions of user-generated communications over SM have been considered sufficient for extraction of self-admission of various use of xenobiotics via NER and hybrid methods.[130]

To develop a compendium of all possible xenobiotics that users could have been exposed to, an initial synset consisting of over 2,000 pharmaceutical and over-the-counter/supplement xenobiotic substances (with over 7,000 related brand and compounded preparation names) was retrieved from the United States Food & Drug Administration (US FDA)'s *Orange Book*[131] master database file.

The file was then parsed to separate multi-compound brands (e.g., the product MUCINEX was originally listed as containing "dextromethorphan and guaifenesin"; the entry was converted to *MUCINEX,dextromethorphan* and *MUCINEX,guaifenesin*). Anion references in salt compounds were removed from generic and brand names (e.g., *paroxetine mesylate* became simply *paroxetine*; *duloxetine hydrochloride* became simply *duloxetine;* etc.). These modifications truncated the original *Orange Book* file into having 1,479 unique pharmaceutical xenobiotic compounds named, with 4,467 associated brand name pairs.

Most important is the fact that the synsets were enriched with every misspelling and abbreviation found during manual annotation of the training set. Over 100 unique drug misspellings were discovered via annotation and appropriately mapped to the synset. In addition, 30 observed instances of foreign brand name usage (virtually all United Kingdom-market names) were added.

Nonetheless, even with the enrichment of synsets with user-generated misspellings and abbreviations, NER via the synsets still contains one major weakness: It can only find the mention of a xenobiotic substance and cannot affirm whether the user claimed to have used it.

Therefore, a risk of elevated type I error (false positive) may exist, whereby users are falsely assumed to have been taking a substance simply because they mention it. Similar risks have already been identified in the processing of clinical narratives (the free-text of electronic medical records and published clinical case study text).[132]

True positive: I have been on aza for 6 months now False positive A (observer): I heard aza causes agranulocytosis False positive B (negation): I never took aza

Figure 7. NER True and False Positive Examples

As seen in Figure 7 above, term *aza* is correctly used in the NER synset to refer to azathioprine, but false positives occur because (A) a communication was simply mentioning the xenobiotic substance in passing or because (B) the communication indicates denial of the user taking the substance.

Therefore, in the previous steps, the entire phrase (complete with subject, verb, and object) was recorded, as the subject and verb content are important to determine whether the user was actually taking the substance in question. The subject and verb were taken as prefix and the actual manner in which the compound was referred to taken as the target. Prefixes recorded both the experiencer and temporal aspects of the xenobiotic's mention in a single element. Examples of actual prefixes discovered are shown in 4.1.2 (Results, Prefixes). Furthermore, suffixes, of few of which were discovered in annotation, were added to create an affirmative (albeit experiencer-ambiguous) class.

Finally, increased Type II error can occur during literal matching via user-generated misspellings (other than those encountered in the training set) missed by the synsets. Therefore, a basic local alignment search (BLAS; BLAST)[129], similar to that used for alignment of genetic sequences, is utilized in order to match non-identical but likely similar strings. For each synset member, if the communication fails to contain it literally, the member is then compared pairwise across every possible frame through the entire communication.

The prefixes and suffixes were integrated into NER by enhancing an existing Java Virtual Machine (JVM) synset detector that was the earlier work of the author and colleagues (AK; JJ).[133] Prefixes and suffixes are searched for in a 45 character geographical neighborhood of the synset member. Therefore, the first phase of search is pure named entity recognition, and further search steps are a hybrid of NER and rules-based search (RBS).

Finally, as indicated in the following diagram, a step to implement BLAS (Fuzzy) search is carried out in case of a synset member not being found literally. Such instance occurs a majority of the time, because most communications do not contain many individual synset members.

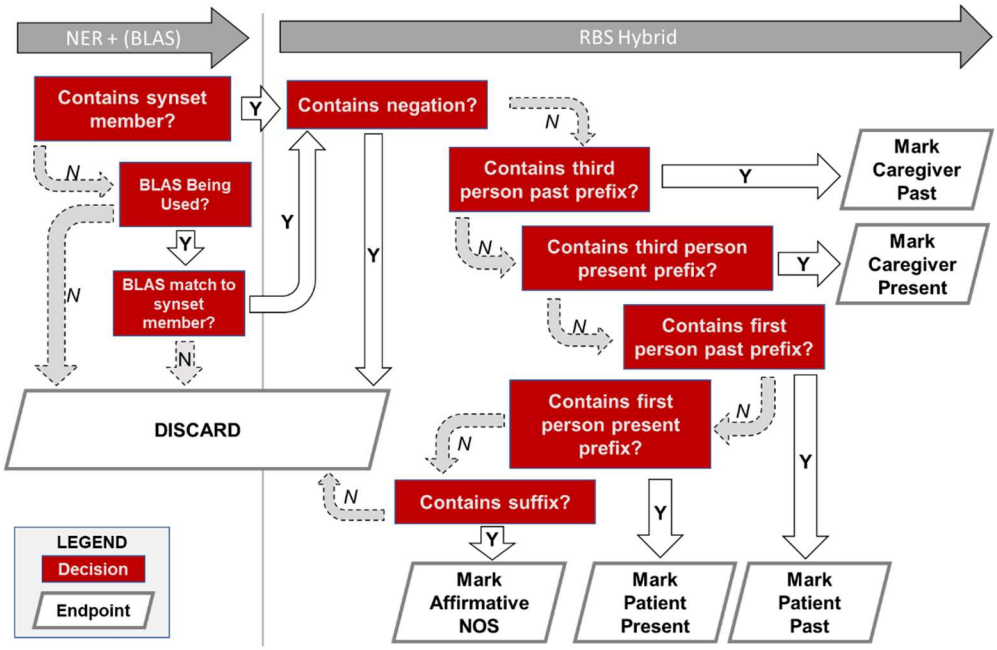

Figure 8. NLP Hybrid Methodology Workflow

The algorithm, depicted in Figure 8, was then run on the testing set of 40 user narratives. Precision and recall were judged on correctly identified user-xenobiotic pairs. Although tense and specific experiencer are recorded by the algorithm, they are not analyzed here as they are not significant to the aims of the research at hand and serve only to classify the communication as evidencing xenobiotic intake and not simply observation of chemical consumption.

With testing set analysis successful, a corpus-wide search of all 1,052 user narratives is performed to describe and quantify xenobiotic usage in this virtual cohort. This search was performed by executing the aforementioned algorithm across the corpus of over 18,000 group communications.

5.3 Results

The 73 narratives in the testing set contained a total of approximately 2,000 posts and comments that were subject to annotation. 599 pairs of users referring to xenobiotics were noted, of which 461 pairs indicated the user (or patient user was caregiver for) actually taking the xenobiotic in question.

One very significant change to the structure of the experiment was effected due to what was observed during training set annotations: The two common AIH treatment compounds prednisone and prednisolone were both determined to be frequently referred to as *pred* by the patients, and had to be merged into one ambiguous term because it was often impossible to differentiate them on the basis of lexical expression, the only available hint. The following metrics in Table 7 describe the different kinds of medication spellings (assessed post-adjudication) used by patients, caregivers, and observers:

Table 7. Types of Xenobiotic/Drug Spellings Utilized

Spelling type		Patient	Caregiver	Observer	Total
Correct	Brand	125 (30.6%)	10 (15.9%)	26 (21.3%)	161 (27.2%)
	Generic	150 (36.8%)	32 (50.8%)	60 (49.2%)	242 (40.8%)
	Drug Class	20 (4.9%)	1 (1.6%)	9 (7.4%)	30 (5.1%)
Clinical Abbreviation	Brand	1 (0.3%)	1 (1.6%)	0 (0.0%)	2 (0.3%)
	Generic	69 (16.9%)	14 (22.2%)	13 (10.7%)	96 (16.2%)
Misspelling	Brand	19 (4.7%)	0 (0.0%)	7 (5.7%)	26 (4.4%)
	Generic	21 (5.2%)	5 (7.9%)	6 (4.9%)	32 (5.4%)
Slang		3 (0.7%)	0 (0.0%)	1 (0.8%)	4 (0.7%)

Some common prefixes are noted below in Table 8. These prefixes hint the machine that the quoted user is actually taking (or caring for someone taking) a xenobiotic, and is neither casually mentioning the product or outright denying usage of it.

Table 8. Drug Intake Prefixes

Experiencer	**Temporal**	**Prefix(es)**	
Patient	Present	Am taking	Take
		Am on	Still on
	Past	Took	Was taking
	Negated	Was never on	Never had
Caregiver	Present	She is on	Husband takes
	Past	He was on	Wife took
	Negated	Daughter never took	She never had

Prefixes for observers (casual mentions) were not created because the exclusion of the experiencer prefixes in rules-based search would automatically create the assumption that the xenobiotic mention was only an observation.

Natural language processing (NLP) was performed on the testing set (35 users who made approximately 500 communications) utilizing named entity recognition hybridized with RBS both with and without and BLAS. Pairs of users and xenobiotic intake affirmations were compared between the annotated *gold standard* set and the result set derived from NLP. Results for recall, precision, and F-score were noted to be as shown in Table 9.

Table 9. Performance of the NER Algorithms

Protocol	**Precision**	**Recall**	**F-score**
NER + RBS (No BLAS; Literal Only)	0.707	0.693	0.700
NER + RBS + BLAS (Literal + Fuzzy)	0.704	0.769	0.735

In order to benchmark the methodology against more comparable studies (i.e., ones that only studied one variable), the precision, recall, and F-score measures were calculated for the four most popular xenobiotics (azathioprine, prednisone/prednisolone, mycophenolate, and budesonide) seen in the annotated gold standard set; benchmark results are shown in Table 10.

Table 10. Performance over Common Xenobiotics

Xenobiotic/Drug	Precision	Recall	F-Score
(Taking any drug at all)	1.000	0.958	0.979
Azathioprine	0.769	1.000	0.870
Prednisone & Prednisolone	0.923	0.800	0.857
Mycophenolate	1.000	0.667	0.800
Budesonide	0.800	1.000	0.889

Error analysis, the qualitative judgment of false positive and negatives, was conducted in order to observe the most common reasons for NLP either failing to match a taken xenobiotic or mistakenly matching one that was not taken. All instances of false results were a result of one of the following conditions:

- (Ambiguity) *Prednisone* and *prednisolone* could not be disambiguated because many users conflated the two medications simply as *pred*. Therefore, users could only be tagged with the fact they were using one or both medications with no further specification. Note that NLP quality metrics exclude this limitation.
- (False Negative) In one user instance, the user was judged by both annotators to have implied intake of multiple xenobiotics due to having mentioned them all in a row without having used any term explicitly indicating intake. The annotators believed that the users were giving out personal medication lists in response to an inquiry about their medication regimen; the current NLP does not detect when multiple xenobiotics are mentioned in a row.
- (False Negatives) Two users with exceptionally high tendencies to misspell words misspelled drug product names to a BLAS similarity score of < 0.50, evading tagging.
- Prednicarbate (False Positives)
    - Resembles prednisone to a BLAS similarity score of 0.75, resulting in false matches when a user says *prednisone* literally.

- Sold as DERMATOP, which closely enough resembles *dermatologist* for a false match to be returned when a user mentions *dermatologist* preceded by words indicating taking a xenobiotic.

After validation of the testing set, the corpus of all 1,052 users was searched for medication usage via the NER + RBS hybrid algorithm. 608/1,052 (57.79%) of users mentioned intake of at least one studied xenobiotic. The most popular xenobiotics, by users claiming having taken them, were therefore revealed and are shown in Table 11.

Table 11. Sample: Xenobiotics Taken by at least Twenty Users

Rank	Xenobiotic	N	% Users	% Users Reporting
	(Reported any xenobiotic intake)	574	54.6%	100.0%
1	Azathioprine	339	32.2%	59.1%
2	Prednisone & Prednisolone Combined	285	27.1%	49.7%
3	Mycophenolate (All Forms)	99	9.4%	17.3%
4	Budesonide	89	8.5%	15.5%
5	Ursodiol	59	5.6%	10.3%
6	6-Mercaptopurine (6-MP)	48	4.6%	8.4%
7	Acetaminophen (Paracetamol)	44	4.2%	7.7%
8	Vitamin D (All Forms)	42	4.0%	7.3%
9	Tacrolimus (Fujimycin)	36	3.4%	6.3%
10	Calcium	31	3.0%	5.4%
11	Diuretic (NOS)	27	2.6%	4.7%
12	Magnesium	23	2.2%	4.0%
13 (T)	Amoxicillin	20	1.9%	3.5%
13 (T)	Ibuprofen	20	1.9%	3.5%

5.4 Discussion & Conclusions

Percentage-wise differences were discovered in how users as patients, caregivers, and observers of xenobiotics referred to these products. The communication group most likely to use correctly spelled brand names was observed to be those communicating as patients. Generic drug names were most likely to be used by those acting as caregivers, while correctly spelled drug classes were seen disproportionately frequently among communications that mentioned the xenobiotic only in passing.

Clinically acceptable abbreviations of brand names were relatively rare, and more common for generic names. Misspellings, overall, were surprisingly infrequent (9.8% of all mentions), raising speculation that group members could be affected by a *white coat syndrome* and use proper medical terminology due to AC's presence. Alternately, it is possible, but likewise not confirmed, that they may be what are considered *expert users* due to their extensive experiences with the medical system. Furthermore, it is noted that slang usage was less common than even misspellings, further hinting at possible backgrounds and motives of the users in the group.

While typical NLP studies prescribe target F-scores in excess of 0.80, such research has only attempted to classify the existence of one variable through a corpus and attempting to do multinomial classification at this level would not be feasible. Many of these studies were also performed solely on electronic medical record (EMR) free text, where brand name usage is less common, misspelling extremely rare, and the mere mention of a medication is sufficient to assume that administration occurred. The closest similar recent study (Klein et al, 2017)[130] demonstrated an F-score of 0.67 for a hybridized approach to detecting multinomial xenobiotic usage across a social media corpus, and the techniques in this dissertation demonstrate a composite F-score of 0.731, higher than the previous effort by those authors.

Even though only 9.8% of all xenobiotic mentions were noted as misspelled in gold standard annotation, adding a fuzzy (BLAS) step to the search algorithm increased recall to 0.766 from 0.693 with only a modest drop in precision. The F-score increased in turn from 0.700 to 0.734. Despite the fact that the addition of BLAS increased the wall clock execution time from 15 minutes to approximately 45 minutes to process the entire corpus, BLAS is still considered valuable for increasing recall and overall quality due to the increase in F-score.

Finally, a comparison to single-compound detection studies can be performed by reducing the results to a univariate format, comparing precision and recall using only singular xenobiotics. Detection performances of azathioprine, budesonide, mycophenolate, and prednisone/prednisolone usages were above 0.800.

With a sufficiently valid methodology developed to analyze xenobiotic usage over social media, the corpus of all 1,052 group users could be analyzed. The most common xenobiotics indicated as taken were therapeutics for AIH and its most common sequelae.

Hydroxychloroquine is an anti-malarial drug that is also used to modify the progression of a common autoimmune comorbidity (rheumatoid arthritis) [11] of AIH. Lactulose is typically used to relieve the etiological factors behind hepatic encephalopathy (HE), a known AIH sequela. Other xenobiotics estimated to be consumed, such as acetaminophen, ibuprofen, and amoxicillin, are very commonly used in the general population. Furthermore, amoxicillin is commonly used by the general population; it and other antibiotics, especially in combination preparations with clavulanic acid have been linked by researchers to hepatic injury. [17] Finally, the role of pain in AIH symptomatology may be suggested by the high rankings of tramadol (an opioid pain reliever) and gabapentin (a neuropathic pain drug) intake, in addition to claimed intake of more common analgesics.

Furthermore, common vitamin/mineral supplements mentioned as taken included vitamin D (all forms), magnesium, and calcium products. In addition, the less commonly used supplement melatonin, used commonly as a sleep aid, was mentioned often as taken. Intake of

these substances might suggest treatment of untoward AIH (or comorbidity) symptom treatment and possibly polypharmacy of AIH medication side effects, although more data, in addition to consults with clinicians, are required to elucidate this relationship.

One significant issue discovered during gold standard annotation was that the compounds prednisone and prednisolone could not be differentiated in automated search due to many users conflating them by using the term *pred*. It is possible that contextually searching the remainder of the user's communications can reveal which xenobiotic the user is actually referring to, but the NLP method does not analyze user postings in such a manner.

Furthermore, a broader limitation is noticed: The information derived faces the same problems electronic medical record (EMR)-derived medication lists. A primary issue was incomplete reporting, noted in that only 337 out of 1,052 group users reported intake of azathioprine, a medication used to treat a far higher proportion of AIH patients. Furthermore, the overall figure for users having claimed to have taken any xenobiotic at all was 574/1,052 (54.6%) which is also likely to be an underestimate of the real figure. The other significant issue shared with EMR-derived research was the inability to reliably judge when a patient (user) stopped or started a medication. While past and current medication usage are differentiated with the algorithm in order to differentiate patients/caregivers from mere observers, the tense-based portion of classification has not been subject to quality assurance. Future research therefore should involve more detailed study into temporal attributes of medication usage in this and similar cohorts.

Finally, the algorithm utilized is computationally inefficient due to its use of literal text matching and moreover, lack of BLAS time optimization. Wall clock time of scanning all 18,000+ communications was under 15 minutes without BLAS but increased to over 2 hours when BLAS was added. Future research must consider ways in which the BLAS portion can be optimized for more efficient and expedient throughput.

Despite the limitations encountered, the proposed executed methodologies are valid for detecting xenobiotic usage and thus performing surveillance on a virtual cohort of over 1,000 users of an autoimmune hepatitis (AIH)-related online support group. The reliability has surpassed that of the most recent similar study. [130] As per prefix detection in section (4.1.3), combined with synsets involving brand names and a BLAS-based correction for potential misspellings, the research at hand has also created an effective *folksonomy* to classify xenobiotic usage via consumer- (patient-) generated corpora of communications. With the research that has been generated, the potential exists for the examination of any user-generated text corpus, in particular one relating to health, for intake of xenobiotics and drugs.

CHAPTER 6. DETECTING CLINICAL FACTORS EXPRESSED OVER THE AIH GROUP

6.1 Introduction

Drug (xenobiotic) intake, as analyzed in the previous chapter, is considered only one clinical concern in Autoimmune Hepatitis (AIH). AIH is shown to have significant gastrointestinal sequelae of hepatic destruction and has been associated with highly associated gastrointestinal and systemic autoimmune comorbidities. [9, 10, 33-36] Frequently experienced sequelae include, edema, hepatic cirrhosis, and hepatic encephalopathy. AIH may co-occur with and has the potential to be exacerbated by other liver-related diseases, including Diabetes Mellitus (DM). [37-39] Other signs and symptoms are likely due to these sequelae as well as the adverse effects of treatments: Changes in weight, impaired white blood cell (WBC) function, and severe pain in and around affected regions of the body are frequently noted .[40] is therefore of significant interest to classify AIH patients by sequelae and symptoms experienced.

Symptoms may, per clinician judgement, also be due to adverse drug effects (ADEs; e.g., *side effects*). Corticosteroids, a mainstay of AIH treatment, are documented to have heavy burden of ADEs, some of which (such as edema, weight changes, and impaired WBC function[40]) are common to original disease symptoms and were found in one study to in fact decrease AIH patient quality of life.[41] Such drugs and regimens are known to aggravate the paralyzing fatigue and psychological risks[42] already observed in untreated disease. As noted in the next section, pain is a common indirect sequela of AIH treatment.

Regardless of alleged cause, the important symptom of pain (physical and psychological), along with fatigue, is a major concern in patient quality of life (QOL). Pain has been noted at elevated rates in the AIH population; mycophenolate[44, 45] usage may be one correlate. Corticosteroid use is known to correlate with pancreatitis,[47] soft tissue injuries[42] and

osteonecrosis-induced bone fractures,[46] which can all be painful. Autoimmune comorbidities of AIH also include painful syndromes of arthritis.[11]

Pain in the form of psychological distress (and potentially psychosomatic physical pain) is seen in AIH, potentially at an elevated rate compared to control populations.[48-50] Mental depression is in particular considered common in AIH[49, 50] furthermore, immunosuppressant therapies (in particular, corticosteroids[51]) are associated with similar adverse neuropsychiatric effects. Furthermore, the mental trauma incurred due to the exceptional symptom burden and resultant decreased quality of life contribute to psychological distress.[70]

Fatigue, a biopsychological symptom, is also noted but the etiology considered cryptic; it has been attributed by some researchers as an ADE.[42, 52, 53] Cognitive impairment is observed and has been attributed to hepatic encephalopathy, a condition where the brain is overloaded with toxic metabolic byproducts due to the liver failing.[32, 54]

Batteries of liver function tests (LFTs) are conducted regularly on AIH patients in order to gauge biological disease severity.[55] In addition, various auto-antibodies are assayed despite the fact that they are not always considered diagnostic of AIH.[56]

For all of these signs, symptoms, comorbidities, and even test results, social media can be amenable to syndromic surveillance (SS), a public health-based technique of recording symptoms without observing patients directly. Social media has been used for gathering reference data for epidemiologic studies[134] and environmental and sociodemographic factors into infectious disease prediction models.[135] In fact, the author's earlier[114] and forthcoming[43] research ha suggested that signs, symptoms, comorbidities, and serum monitoring (lab) results are frequently discussed on the AIH group in question.

Therefore, the purpose of this phase of dissertation research is to determine the feasibility of conducting online surveillance of:

- Signs & Symptoms
- Diagnosed Comorbidities
- Laboratory test results

…associated with autoimmune hepatitis (AIH) over an AIH-oriented social media support group.

6.2 Methodology

Initially, a qualitative analysis on a testing set of the entire communications of 73 users (over 1,500 communications) was performed in order to determine the manner in which signs, symptoms, diagnosed comorbidities, and lab test results were expressed by the online cohort in question.

Annotations were standardized using the Systematic Nomenclature of Medicine (SNOMED-CT). Each annotation was mapped to a preferred SNOMED-CT *finding*-class term with 264 SNOMED finding-class concepts discovered. Afterwards, the preferred finding-class terms were clustered by mutual agreement of both annotators into 49 different classes of findings.

Appendix IV contains a list of all SNOMED concepts whose experience was discovered during training set annotation. In addition, it contains the mapping of these concepts to the authors' 49 chosen higher-level concepts.

After annotation, literal mentions of these factors were also qualitatively clustered by semantic regular expression traits and the following semantic clusters (not identical to the 49 SNOMED clusters) of mentions determined as shown in Table 12.

Table 12. Regular Expressions by Algorithm

Factor	Expression (Any Order)	Expression List In
Pain-related sign/symptom	Body Part (AND) Term Indicating Pain. (Sequential regular expressions; please see below figure)	Appendix V
Non-pain-related sign/symptom	(Experiencer Term) (AND) Term Indicating Sign or Symptom	Appendix VI
Diagnosed Comorbidity	(Experiencer Term) (AND) Term Indicating Comorbidity	Appendix VI
Lab Test Result	Name of Test (AND) Indicator of Value	Appendix VII

Each cluster is characterized by a regular expression, a set of features that are required in order to indicate the user was experiencing the problem in question. Note that the order of elements in each expression is not binding, but the presence of all elements is.

Therefore, the most suitable way to search the wider corpus of postings was determined to be via three separate algorithms, with each algorithm using its own expression content:

- Pain
- Diagnosed comorbidity & non-pain-related signs and symptoms
- Lab test results

As per the cardinal expression (#1) detailed in (6.3.1), evidence of experiencing pain was searched by the hybrid named entity recognition (NER) and rules-based search (RBS) algorithm described in Figure 9.

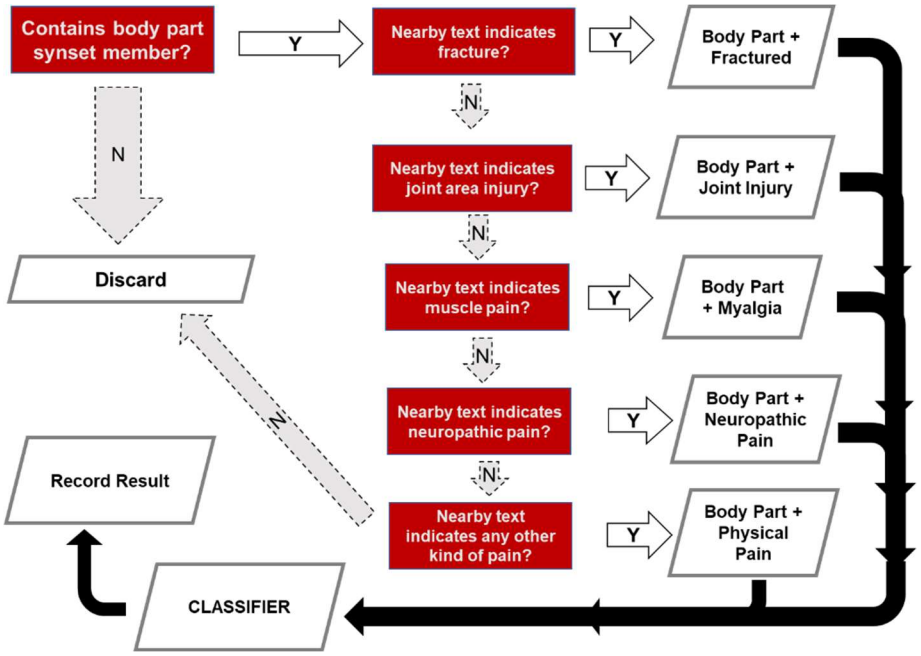

Figure 9. Sequential Pain Detection Algorithm

Depending on the body part and type of pain, the pain classification was reduced from (10 * 5) = 50 to 10 classifications by clinical judgment of the second annotator, subject to approval by the primary author. The classifications and logic thereof are seen in Figure 10.

		Type of Pain or Injury				
		Neuropathy (Burning, Tingling, Parasthesia)	Fracture (Breaking a bone)	Soft Tissue Injury (Sprains, Strains, Bruises, Dislocations)	Myalgia (Cramps, Soreness, Fibro)	Physical Pain NOS (Hurting, Aching)
Body Part/Region	Lower Limb (Excl. Joints)	Neuropathic Pain (247398009)	Fracture of Bone (125605004)	Soft Tissue Injury (282026002)	Muscle Pain (68962001)	(Physical) Pain (22253000)
	Upper Limb (Excl. Joints)					
	Back/Spine					
	Dental/Oral	Painful Mouth (102616008) [NB3]			Painful Mouth (102616008) [NB3]	
	Thorax	Neuropathic Pain (247398009)			Muscle Pain (68962001)	(Physical) Pain (22253000)
	Head				Headache (25064002)	
	Abdomen (Including Quadrants)		*Not Observed*		Abdominal Pain (21522001) [NB4]	
	Bone (Any)	*Not Observed*	Fracture of Bone (125605004)		Bone Pain (12584003)	
	Joint (Any)	Arthralgia (57676002) [NB2]			Arthralgia (57676002)	
	Pain but no body part	Neuropathic Pain (247398009)			Muscle Pain (68962001)	(Physical) Pain (22253000)
	NOTES	[NB1]	*Burning/tingling in abdomen typically due to organomegaly associated compression*			
		[NB2]	*Burning sensation in joints typically due to arthritis, so classified as arthralgia*			
		[NB3]	*Oral myalgia classified as mouth pain due to soreness typically being not muscle related*			
		[NB4]	*Abdominal myalgia classified as abdominal pain due to cramps*			

Figure 10. Higher Level Classification of Pain and Injury

The algorithms for non-pain-related signs and symptoms and diagnosed comorbidities are far simpler and therefore not diagrammed. The analogous classification table is available in Appendix IV.

The following lab test results, important in the monitoring of the disease, were sought:

- Liver Function Tests (LFTs)
    - Bilirubin
    - Alanine transaminase (ALT)
    - Aspartate transaminase (AST)
    - Alkaline phosphatase (ALP)
    - Gamma-glutamyl transpeptidase (GGT)

- Autoimmune antibody (AAB) titers
    - ANA
    - PANCA
    - LKM
- White blood cell (WBC) titers
    - Lymphocytes
    - Neutrophils
- Red blood cell (RBC) titer

Based on patient self-report as test result high, low, abnormal, or normal, the annotators agreed upon higher-level classifications for test results as schematized in Figure 11.

		Test Type			
		Liver Function (LFT)	Autoantibody (AAB)	Red Blood Cell (RBC)	White Blood Cell (WBC)
Level	High	LFT Abnormal / AIH Flare (707724006 / 408335007.FLARE)	Autoantibody Titer Positive (165878000)	RBC Normal (165421004)	WBC Normal (165507003)
	Low	LFT Normal / Serum AIH Remission (166602006 / 408335007.STAB)	Autoantibody Titer Negative (165878000)	Anemia (271737000)	Leukopenia (NOS) (419188005)
	Abnormal	LFT Abnormal / AIH Flare (707724006 / 408335007.FLARE)	Autoantibody Titer Positive (165878000)		
	Normal	LFT Normal / Serum AIH Remission (166602006 / 408335007.STAB)	Autoantibody Titer Negative (165878000)	RBC Normal (165421004)	WBC Normal (165507003)

Figure 11. Higher Level Classification of Lab Test Results

To validate the findings, a test set of 35 users was then analyzed for signs, symptoms, pain signs, pain symptoms, comorbidities, and laboratory test results. The aforementioned algorithms were then run over this new corpus of users and measures for sensitivity, specificity, and F-score calculated.

6.3 Results

Performance results varied depending upon the algorithm, granularity of detection sought, and specific problem to be detected. Relevant results are detailed in the below tables (Tables 13 and 14).

Table 13. Algorithm Performance: Pain and Injury

Note: Only combinations that evidenced a gold standard N of >=5 are discussed.

Observation Granularity	Variable Granularity	Precision	Recall	F-score	(N +)
User (35)	Any type of pain or injury	0.778	1.000	0.875	15
Communi-cation (277)	Any type of pain or injury	0.722	0.813	0.765	135
User (35)	Correct type of pain or injury (20 types)	0.667	0.667	0.667	15
	Arthralgia only	0.800	0.889	0.842	11
	Physical Pain NOS only	0.625	0.625	0.625	8
	Abdominal pain only	0.600	0.600	0.600	5

The next table represents the detection of non-pain signs, non-pain symptoms, and diagnosed comorbidities.

Table 14. Algorithm Performance: Signs, Symptoms, Comorbidities

Variable	SNOMED-CT Code	Precision	Recall	F-score	(N +) (/35)
Correct sign, symptom, comorbidity type	N/A	0.694	0.767	0.729	31
Cirrhosis	19943007	0.909	1.000	0.952	10
Comorbid autoimmune disorder	85828009	0.889	0.800	0.842	10
Fatigue	84229001	0.600	0.667	0.632	10
Infectious disease	40733004	0.667	0.667	0.667	9
Mood issue	271596009	0.643	1.000	0.783	9
Gastrointestinal upset	162059005	0.667	0.800	0.727	7
Fibromyalgia	203082005	1.000	1.000	1.000	5
Cognitive Impairment NOS	386806002	1.000	0.750	0.857	4
Disorder of Pancreas	3855007	1.000	0.750	0.857	4
Jaundice	18165001	0.667	1.000	0.800	4
Note: Only entities that achieved a rate of at least 4 gold standard testing set positives are shown here. The observation granularity is always by user (never by communication).					

Finally, results pertaining to the detection of specific aspects of laboratory results are shown in Table 15.

Table 15. Algorithm Performance: Lab Test Results

Variable Granularity	Permu-tations	Precision	Recall	F-score	(N +) (/35)
Correct test performed on user	3	0.818	1.000	0.900	27
Correct test with correct result	12	0.605	0.676	0.639	27
Liver function test (LFT) with correct result	4	0.692	0.720	0.706	19
Autoantibody (AAB) titer with correct result	4	0.429	0.600	0.500	5
White blood cell count with correct result	4	0.400	0.500	0.444	4
Note: The observation granularity is always by user (never by communication).					

Common patterns of errors were elucidated that hampered precise detection of the studied factors. The following errors were known to occur at least twice upon further examination:

- Psychological pain/mental distress would be referred to with terms such as *it hurts, in pain, hurts me,* and other false positive n-grams that the algorithm coded as physical pain.

- The abdominal pain dictionary included various abdominal organs. Users often mentioned abdominal organ damage followed by pain in another part of the body. In this case, the algorithm tagged users as having pain both in the abdomen (false positive) and the other body part (true positive).
- In detection of AAB titer self-reports, positive and negative readings were often conflated. Two instances were found where the user mentioned one antibody as positive and one as negative within the same sentence; the combination of antibody readings into one variable made these users default to self-reporting negative AAB titer status.

The algorithm was then run on the entire corpus in order to gain population estimates of patient self-reports of these signs, symptoms, comorbidities, and lab test results. Only aspects that attained a test set (gold standard) N of at least 4 are listed in summary Tables 16, 17, and 18, below.

Table 16. Pain & Injury over the Wider Corpus

Pain Type, Classified	SNOMED-CT Code	N Users (/1052)	% Users	% Report Users
Any type of pain or injury	N/A	365	34.7%	100.0%
Physical Pain NOS	22253000	205	19.5%	56.2%
Arthralgia	57676002	137	13.0%	37.5%
Abdominal Pain	21522001	96	9.1%	26.3%

Table 17. Signs, Symptoms, and Comorbidities over the Wider Corpus

Variable	SNOMED-CT Code	N Users (/1052)	% Users	% Report Users
Any sign, symptom, comorbidity	N/A	687	65.3%	100.0%
Mood issue	271596009	249	23.7%	36.2%
Infectious disease	40733004	222	21.1%	32.3%
Fatigue	84229001	219	20.8%	31.9%
Comorbid autoimmune disorder	85828009	137	13.0%	19.9%
Cirrhosis	19943007	136	12.9%	19.8%
Gastrointestinal upset	162059005	112	10.6%	16.3%
Cognitive Impairment NOS	386806002	112	10.6%	16.3%
Jaundice	18165001	82	7.8%	11.9%
Fibromyalgia	203082005	42	4.0%	6.1%
Disorder of Pancreas	3855007	22	2.1%	3.2%

Table 18. Lab Test Results over the Wider Corpus

Variable	N Users (/1052)	% Users	% Report Users
Liver function test (LFT) result reported	386	36.6%	100.0%
Normal LFT (AIH Serum Stabilized)	203	19.3%	52.6%
Abnormal LFT (AIH Flare)	323	30.7%	83.7%
Only Normal LFT Reported	63	6.0%	16.3%
Only Abnormal LFT Reported	183	17.4%	47.4%
Both Normal & Abnormal LFTs Reported	140	13.3%	36.3%

6.4 Discussion

Our research has shown evidence for the feasibility, via targeted natural language processing (NLP), of the extraction of various clinical factors self-reported to have been experienced by patients of autoimmune hepatitis (AIH). Multiple algorithms allowed for the automated discovery of self-reported signs, symptoms, comorbidities, pain-related issues, and disease monitoring results across a corpus of 1,052 members of an AIH-oriented support group.

In terms of detecting pain symptoms and injury, the algorithm performed best in detecting self-reported arthralgia (i.e., joint pain), with an *F*-score of 0.842. Detection of expressed general physical pain (not otherwise specified as to nature or location) [136] and abdominal pain were satisfactory, with *F*-scores of 0.625 and 0.600, respectively. Furthermore, the detection of any user (or in fact any communication) evidencing any sort of pain or injury was reliable at $F = 0.875$ for users and $F = 0.765$ for individual communications.

Set upon the wider corpus of users, the pain detection algorithm demonstrated that unspecified physical pain was most commonly self-reported (19.5% of all users, or 56.2% of pain-reporting users). Arthralgia was also noted by 13.0% of all users (37.5% of pain-reporting users). Unspecified pain is expected to be most common due to its physical generality; on the other hand, arthralgia is more remarkable because is not as common in the general population and is still nascently being researched as an AIH symptom that may be due to autoimmune-mediated joint destruction.[136] The reporting rate of abdominal pain was at 9.1% of all users (26.3% of pain-reporting users) and therefore lower than the self-reported rate of arthralgia in this cohort despite the inherently abdominal nature of the disease. The rate of self-reported arthralgia in this group, if corroborated by survey evidence, may support the sparse literature on the topic of arthralgia in AIH, and more research to uncover the relevant links would have to be conducted.

Several non-pain signs, symptoms, and comorbidities were also deemed amenable to automated extractions. The pilot testing analysis showed that mood issues, infectious disease, fatigue, comorbid autoimmune disorders, cirrhosis, gastrointestinal upset, and cognitive impairment were detectable at satisfactory *F*-levels. In addition, for all possible non-pain-related recognized signs, symptoms, and comorbidities, the overall *F*-score for detection was satisfactory at 0.729.

The corpus-wide analysis revealed that 65.3% (N = 687/1052) users expressed one of the sought signs, symptoms, or comorbidities. The most commonly observed self-reported factor pertained to mood issues (which are defined by the annotators and recorded by the algorithm as

all mood-related psychiatric diagnoses along with expressions of profound sadness or grief), which were expressed and/or self-reported by 23.7% of individuals (36.2% of those who expressed any sign, symptom, or comorbidity at all). In addition, infectious disease and fatigue were reported by over 20% of all group users.

Anomalous detection rates were also observed. A diagnosis of cirrhosis, a late-stage complication of AIH, was self-reported by 12.9% of all group users. In comparison, researchers estimate that 25% of AIH patients have cirrhosis *at initial disease presentation* (with a much higher lifetime prevalence)[31], suggesting that cirrhosis may be underreported by users in the cohort currently at study. Gastrointestinal upset may also be underreported (10.6% of all users; 16.3% of all reporting users), especially viewed in comparison to cognitive impairment, which was self-reported at an identical frequency.

Of the laboratory test results, only self-reported liver function tests (LFTs) were demonstrated to be reliably ascertainable; this detection feature allowed for partial classification of users claiming to have experienced disease flares and/or remissions.

The limitations inherent in this research may form a basis for future research that should be conducted on the subject. A chief limitation that occurred in classification was the fact that although SNOMED-CT contains unique concept mappings for clinical entities, these entities are technically subject to type constriction, with entities typically typed as sign/symptom, finding, entity, and diagnosis.

The prime limitation was noted with regards to non-pain signs, symptoms, and comorbidities. Specifically, although the F-score measure of reliability was satisfactory (0.729), there existed a limited span of covered issues. Similarly, the small size of the testing set (N = 35 users) limited the overall study to pilot status.

While this research demonstrates the feasibility of extracting certain clinical signs and factors from a cohort of people affected with AIH, it cannot demonstrate the correlations these factors have with each other or with other clinical factors such as medication intake. Associative

analyses will be an important part of future research and while they will not definitively create relationships with treatments and effects, can still help inform clinical workflows.

CHAPTER 7. DETECTING CONTEXTUAL FACTORS EXPERIENCED BY AIH PATIENTS

7.1 Introduction & Background

While I have already well-entertained the spectrum of factors already ascertainable in the AIH clinic, it is necessary to examine the AIH patient beyond the traditional clinical knowledge environment. *Contextual factors* for purposes of this discussion are non-clinical factors present in any individual or population thereof. Contextual factors include (but are not limited to) demographics, quality of life (QOL), physical and social environments, and lifestyle-related factors. At this point, I speculate and research the prospect that social media can be suitable for the potential of ascertaining non-clinical factors of a body of AIH patients.

A two-search systematic review and synthesis of the literature, detailed in the overall background (Chapter 2) of this dissertation, discovered multiple categories (i.e., themes) of contextual factors that have been researched in AIH patients. These themes were diet-related factors, treatment noncompliance (patient-initiated treatment withdrawal), physical environment, psychological quality of life (QOL), recreational substance usage, and research advocacy (ADVOC) for further contextual factors research.

The existing themes from the literature review form part of the framework needed for classification of contextual factors as potentially observed on the AIH-related Facebook™ group that is the subject of this dissertation's research. Furthermore, a *thesaurus*-based format of classification as prescribed by Holden et al[67] is utilized, where each category (theme) and each component (sub-theme or factor) that makes it up are also explicitly and unambiguously defined.

I therefore wish to investigate the creation of tools that would be required for the automated *tagging* of user-generated communications in a fashion that is relevant to improving the user's health. Therefore, customizing an algorithm with which multifaceted user text corpora may be classified is of essence to achieving the goal of tagging potentially health-related user-generated content.

7.2 Methodology

A *top-down* thesaurus-based approach was utilized to synthesize and classify *all* factors, clinical and non-clinical, experienced by patients. Although the focus of the research is with contextual factors, information conveyed by patients in a real-life online support group may also be clinical, be in inquiry for advice, or be of support to another user. Therefore, all facets of all communications (be they clinical, support-related, or non-clinical/contextual) are studied while indexing with the thesaurus.

Annotation was performed with a single annotator (the lead author). Every fragment of every communication was annotated and assigned a thesaurus code. Annotation ceased at saturation (when the annotator had gone through 50 communications in a row without encountering any new concepts at all), resulting in a total of 676 annotated communications split into 882 fragments. Annotation validation was performed by single-blinded re-annotation, the author using the existing list of codes to blindly annotate the corpus once again.

Because the purpose of the research at hand involves automated detection of the contextual factors facing patients (support group users), a computational approach is favored to attempt automated detection of these important facets of patient life. Due to the high number of observed categories (16 top-level domains with over 210 subdomains; see section 7.4.1 for more details), a *machine learning* approach (a form of complex analysis that takes multiple numeric

and nominal factors to characterize observations and predict future ones) was utilized to explore automated detection of contextual factors.

In order to perform machine learning, textual data must be processed into numerical (or Boolean) vectors (with each vector being the values of a set of variables).

For example, let Fruits be a set of variables:

$$Fruits = [Apple, Banana, Tomato, Orange, Lemon]$$

And let Sentence be a sentence:

$$Sentence = [In\ my\ garden\ I\ planted\ four\ tomato\ bushes\ and\ an\ apple\ tree]$$

In this example, neither Sentence nor Fruits are readily computable because they consist of text. However, Sentence can be expressed as a Boolean vector of the variables seen in Fruits:

$$Sentence(Fruits) = [1, 0, 1, 0, 0]$$

…because it contains *tomato* as well as *apple*. Similarly, Sentence can be compared to another vector:

$$Animals = [Cow, Human, Pig, Sheep]$$

In that case,

$$Sentence(Animals) = [0, 0, 0, 0, 0]$$

And is fully represented by

$$Sentence = [Fruits(1, 0, 1, 0, 0);\ Animals(0, 0, 0, 0)]$$

As well as in relation with respect to Sentence's length:

$$Sentence = [Fruits(0.083, 0.0, 0.083, 0.0, 0.0);\ Animals(0.0, 0.0\ 0.0, 0.0)]$$

As well as in relation with respect to the number of Fruits and Animals combined:

$$Sentence = [Fruits(0.111, 0.0, 0.111, 0.0, 0.0);\ Animals(0.0, 0.0, 0.0, 0.0)]$$

Finally, the two vectors expressed as *in relation* may be combined to form a normalized Term Frequency-Inverse Document Frequency (TF-IDF) matrix, which expresses the frequency of words in Sentence in a direct ratio to the relative frequencies of all words in Fruits and Animals. There are multiple ways of calculating TF-IDF, the simplest of which is dividing the probability of the word within a sentence to the probability of the word within the classification array:

$$TFIDF(Sentence) = [Fruits(0.748, 0.0, 0.748, 0.0, 0.0);\ Animals(0.0, 0.0, 0.0, 0.0)]$$

The vectors of Sentence, in turn, can be used to determine what it is about. Because Sentence does match to Fruits to some degree but not to Animals at all, it is safe to say that if a machine had to choose which of those two topics Sentence was about, it would most likely choose Fruit as the topic. Overall, the TF-IDF vector normalization of different groups of text facilitates computational learning by converting text representations into simpler numerical formats.

Preprocessing of text was performed by using a customized dictionary and original Java virtual machine (JVM)-based subroutine to do the following actions:

- Stemming of words (removing common endings such as *-ing, -er,* and also de-pluralizing words)

- Stopword removal (removal of entire words, e.g. *and, or, the,* and special characters such as *amp* and *quot* that do not enhance the semantic meaning of a sentence)
- Standardization
    - Replacement of all web links with *url*
    - Replacement of all AIH pharmaceutical therapy drug names with *aihrx*
    - Replacement of all references to AC with *referencetoAC*
    - Replacement of indicators of first, second, and third person with the standard terms *fpindicator, spindicator, and thpindicator* respectively
    - Replacement of all verbs of being (*be, am, are,* etc.) with *verbofbeing*
    - Replacement of all possessive verbs (*have, got, having,* etc.) with *verbofpossession*

In this particular study, communication fragments were grouped by top level domain as the outcome variable because grouping by thesaurus coding would require in excess of 40,000 communications to be annotated. Weka (University of Hamilton, Waikato, New Zealand)[116], an open-source machine learning platform, was utilized in order to perform analysis on the corpus of communications. Each communication fragment was normalized to standardize common *stopwords* and its content truncated to leave words as stems. Afterwards, TF-IDF conversion was automated by Weka's Filtered Classifier String2WordVector function, which calculated the frequencies of both words and bigrams (two-word chains) relative to corpus frequency for each communication fragment. Finally, the mathematically-transformed communications were analyzed via several methods that all employed Sequential Minimal Optimization (SMO), a transformer that allowed machine learning to be performed a per Figure 12's workflow on data with many possible categories (e.g., the 16 top level domains of contextual factors at study).[137]

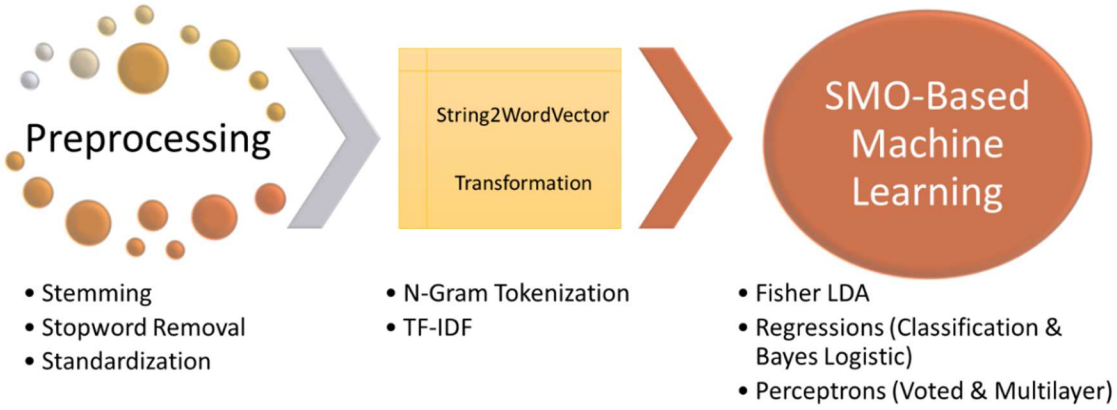

Figure 12. Workflow Schema for ML Contextual Factors Classification

Multiple methods (i.e., algorithms) were used in conjunction with SMO. Each algorithm underwent 10-fold cross-validation to determine its reliability. The methods employed by SMO in detection of contextual factors included:

- Fisher Linear Discriminant Analysis (FLDA): An implementation of principal component analysis that is highly suitable for delineating multiple classes.[138]
- Voted Perceptron (VP): An algorithm that specializes in maximizing the distance (difference) between classes.[139]
- Classification via Regression (CVR): A traditional linear model algorithm that can be tuned to detect multiple classes of different proportions.
- Bayesian Logistic Regression (BLR): Similar to CVR but operates solely on binary trees (and will decompose numbers into said trees)
- Multilayer Perceptron (MLP): An advanced classification algorithm that has shown capability in detecting sparse entities due to its ability to recurse over previously read items.[140] MLP is one of the more common types of neural network (NN) analysis and is somewhat related to VP.

7.3 Results

Qualitative annotation analysis revealed 172 subdomains that were classified under 16 top level domains. Of these subdomains, 156 described actual communications while the remainder served solely as parent domains. The discovered top-level domains are listed below in Table 19 along with their codes and definitions. **A full classified schedule of all 172 subdomain codes and their definitions is included in Appendix VIII.**

Table 19. List of Top-Level Domains with Definitions

Domain	Code	Detailed Definition: The user is discussing…	In Annotated Corpus	Systematic review status
Demographics	DEMO	Commonly elicited social characteristics of an individual	18 (2.0%)	NCF
Physical Environment	ENV	The physical surroundings of a person, including geographical location and natural environment	19 (2.1%)	Y
Family History	FAMHX	Medically-recognized symptoms and diagnoses of blood relatives	15 (1.7%)	NCF
Finances	FIN	Pertaining to monetary instruments, acquirement thereof (e.g., employment) and sureties (e.g., insurance)	20 (2.3%)	Y
Support Group Socializing	GRSOC	Greeting others within context of the group discussion itself	18 (2.0%)	NCF
Healthcare System	HCS	Matters pertaining to the user's interaction with healthcare, including clinic and clinicians/providers. Excludes personal medical story (MEDXS) and insurance (FIN) discussions.	41 (4.6%)	N
Information Sharing	ISHR	Sharing information for academic purposes with no support intended. Indicates advanced knowledge of the subject matter, which is typically health.	30 (3.4%)	N
Medical Stories/ Histories	MEDXS	Medically-recognized elements of health, including diagnoses, medication intake, lab test results, signs, and symptoms. *Communications observed in this domain are primarily detected and classified by the algorithms in Chapter 6 of this work.*	248 (27.9%)	NCF

Domain	Code	Detailed Definition: The user is discussing…	In Annotated Corpus	Systematic review status
Treatment Noncompliance	NCOMP	The act of not adhering to provider-mandated treatments	4 (0.5%)	Y
Quality of Life Factors	QOLS	Psychosocial factors ranging from emotions to coping to abilities to perform the daily activities to one's personal satisfaction	94 (10.6%)	Y
Research Participation	RSPRT	Participating as a subject in medical or social research related to the disorder	4 (0.5%)	N
Social History & Environment	SOC	Factors pertaining to the individuals who physically surround the user. Excludes marital status. Excludes the provision and receiving of online support from other group members.	28 (3.2%)	N
Self-treatment Stories	STX	Sharing stories about personal and self-care treatments, including diet and exercise. These treatments might or might not be at behest of a provider.	44 (4.9%)	Y
Support	SUP	Communications made to request or provide assistance from or to another individual(s). *This category is broken down into subcategories that are entertained and detected further in Chapter 8 of this work.*	278 (31.2%)	NCF
Technology	TECH	The user's relationship with electronic technology	7 (0.8%)	N
Viewpoint	VPT	A personal opinion not necessarily grounded in fact.	22 (2.5%)	NCF
Legend: *Y = Contextual factor that was reflected in the literature review* *N = Contextual factor that was not reflected in the literature review* *NCF = For purposes of this research not considered a contextual factor; some contextual factors (e.g. demographics) easily ascertainable in the clinic are also given this category.*				

Within the annotations that form this thesaurus, it can be remarked on the prevalence of each top-level domain. The most popular top-level topic domains of conversation were by far support (SUP) and medical stories/histories (MEDXS). Together, these top-level domains comprised 526/889 (59.2%) of the annotated corpus. Outside of those two top-level domains,

quality of life factors (QOLS) was the most popular, with 94 communication fragments (10.6% of the corpus). The least popular top-level domains, reflecting the least popular topics of conversation in the sample, were Technology (TECH), Research Participation (RSPRT), and Treatment Noncompliance (NCOMP), each representing less than 1.0% of the corpus. All other top-level domains were discussed at levels representing between 1.0-10.0% of the corpus.

7.3.1 Literature Review Coverage Comparison

The overall numbers and proportions of annotated communication fragments representing each literature review category can therefore be calculated and are expressed in Table 20.

Table 20. Contextual Factors Compared to Literature Presence

Literature review category	# fragments	% fragments
Y (Contextual factor found in communications and in literature review)	181	20.3%
N (Contextual factor found in communications but not in literature review)	10	12.4%
NCF (Not considered a contextual factor)	599	67.3%

7.3.2 Reliability & Performance of Detection Algorithms

The reliability of each algorithm in classifying the identified categories of contextual factors was assessed by *F1*-score; results are noted in Table 21.

Table 21. Algorithm Performance Across Contextual Factor Types

		Algorithm *F1*-Reliability					**Avg. *F1***	**Best *F1***	**Best Algo.**
TLD	**CF?**	**FLDA**	**VP**	**CVR**	**BLR**	**MLP**			
DEMO	N	0.645	0.462	0.514	0.429	0.552	0.520	0.645	FLDA
ENV	**Y**	0.471	0.452	0.387	0.414	0.467	0.438	0.471	FLDA
FAMHX	N	0.348	0.400	0.222	0.381	0.400	0.350	0.400	VP/MLP
FIN	**Y**	0.438	0.500	0.529	0.345	0.160	0.394	0.529	CVR
GRSOC	N	0.414	0.414	0.286	0.320	0.231	0.333	0.414	FLDA/VP
HCS	**Y**	0.386	0.364	0.301	0.370	0.314	0.347	0.386	FLDA
ISHR	**Y**	0.000	0.048	0.195	0.000	0.000	0.049	0.195	CVR
MEDXS	N	0.663	0.638	0.665	0.657	0.652	0.655	0.665	CVR
NCOMP	**Y**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	None
QOLS	**Y**	0.256	0.229	0.295	0.247	0.272	0.260	0.295	CVR
RSPRT	N	0.000	0.000	0.000	0.000	0.000	0.000	0.000	None
SOC	**Y**	0.489	0.508	0.423	0.356	0.545	0.464	0.545	MLP
STX	**Y**	0.301	0.282	0.386	0.290	0.305	0.313	0.386	CVR
SUP	N	0.671	0.640	0.645	0.647	0.656	0.652	0.671	FLDA
TECH	**Y**	0.444	0.444	0.444	0.444	0.444	0.444	0.444	All Tied
VPT	N	0.000	0.085	0.000	0.000	0.000	0.017	0.085	VP
		2/9	0/9	**3/9**	0/9	1/9	**# of Categories Best Algorithm (CF Only)**		
		0/9	**2/9**	1/9	0	1/9	**# of Categories F1 Above 0.500 (CF Only)**		
		0.290	0.285	**0.317**	0.262	0.273	**Wt. Avg (CF Only)**		

Notes:

- *Some top-level domains are conventionally considered contextual factors but because they are typically ascertainable in the clinic are not considered such for purposes of this research.*

Overall, it is noted that optimization of detection of each top-level domain topic was heterogeneous, with different algorithms specializing in detection of different topics. Across top-level domains classified as contextual factors, FLDA and CVR achieved the highest number of domains classified as the best across all algorithms, while CVR maintained the best overall weighted average. VP did not demonstrate top reliability in any specific category yet categorized the highest number (2) of contextual factor top-level domains at F1 >= 0.500. More qualitatively,

despite MLP's modest overall performance, it did by far classify social environment (SOC) the best ($F1 = 0.545$). BLR was, overall, the least-well performing classification algorithm utilized.

7.4 Discussion & Conclusions

The construction of the thesaurus in and of itself is significant in that it (as did Holden et al's research[67]) provides a beginning step to the general classification of contextual issues; this research expands Holden et al's findings to more contextual factors in a rarer disease that affects individuals of all ages. Its top-level categories and many of its specific fine-grained codes can potentially be used to describe contextual factors across other chronic diseases.

Although the core purpose of the research at hand is to attempt delineating contextual factors faced by patients, it can also be commented as to the popularity of other topics of conversation. As far as frequency of discussion top-level domains, it is unsurprising that support (SUP) was the most popular, given that the group at hand is in fact a support group. The status of the second most popular top-level domain topic, MEDXS, may be a bit more unusual, as directly divulging personal medical histories is considered by most to be sensitive. The higher presence of MEDXS, however, validates the fact that in an earlier chapter of this work, *treatment stories* and *treatment side effects* were two of the most popular ML-discovered categories when comparing user vs. user within this corpus. Therefore, it is implied that AIH patients are more likely than what would be intuitively expected to disclose details of their medical histories over an online support group.

Similarly, the majority of topicality via top-level domain in the annotated set was not focused on contextual factors not ascertained in the clinic; only 291 (32.7%) fragments evidenced a topic top-level domain pertaining to such contextual factors. Of these, 181 (20.3% of the corpus) fragments evidenced discussion of a topic identified in the literature review.

The most popular top-level domain contextual factor discussed was quality of life factors (QOLS); this observation is in harmony with the literature's frequent identification of quality of life factors, in particular psychological quality of life, as important to AIH patients' well-being. Self-treatment stories (STX), including those pertaining to diet, also formed a reasonable part of the corpus, aligning with the literature review results that nutrition may be a factor in the pathogenesis of AIH. Furthermore, Finance (FIN), which for these purposes also includes insurance status, was also stated in the literature as a potential AIH-mitigating factor.

Conversely, treatment noncompliance (NCOMP) was repeatedly suggested in the literature but found in only four (0.5%) of corpus communication fragments. This finding may suggest (although cannot confirm) that due to AC's presence in the group, a measure of *white coat* syndrome may occur, where patients do not disclose treatment noncompliance due to disappointing AC.

On the other hand, gaps in literature coverage were also noted. Technology (TECH), social history & environment (SOC), Healthcare System Relationship (HCS), and Information Sharing (ISHR, which may be considered evidence of enhance health literacy), all considered important contextual factors in Holden et al's[67] research, were discovered in the users' communications but not in the literature review. One reason for this discrepancy may be that the original search terms sought *risk factors* for AIH and that current research does not view and therefore does not seek these facets of patient life as potential risk modifiers in AIH.

Perhaps more important than the literature review were the attempts at algorithm-based contextual factor classification. Algorithm results were extremely heterogenous as to which algorithm performed the best in detecting which top-level domain topic of conversation. The below drill-down (Table 22) re-emphasizes individual algorithm performance on each top-level domain.

Table 22. Best Algorithms for each Contextual Factor Category (Theme; Domain)

Top-Level Domain	Best Algorithm	F1 of Best Algorithm
ENV	FLDA	0.471
FIN	CVR	0.529
HCS	FLDA	0.386
ISHR	CVR	0.195
NCOMP	None	0.000
QOLS	CVR	0.295
SOC	MLP	0.545
STX	CVR	0.386
TECH	All Tied	0.444

The overall impression is that algorithm performance tended to be poor-to-modest over detecting contextual factor-related discussion. The best performing algorithm overall (CVR) by weighted score only performed at an *F1* reliability of 0.317. The only top-level domains detectable with reliability of at or above $F1 = 0.500$ were Finance (FIN), detectable by CVR at $F1 = 0.529$ and Social History & Environment (SOC), detectable by MLP at $F1 = 0.545$. It is therefore recommended that any detection algorithms to be deployed in real-world patient health platform situations be employed in tandem, with specific algorithms used to seek specific top-level domain topics of discussion.

A sample of best-performing algorithm (i.e., CVR)-misclassified posts was selected and manually analyzed to determine the nature (and possible reason) for misclassification. The following common misclassifications were noted, and the potential reasons speculated:

- ISHR (Information Sharing) misclassified as SUP (Support): Many support posts evidenced competent use of medical terminology and thus confounded the algorithm into assuming that such communications were a part of SUP, the most common category.
- ISHR misclassified as MEDXS (Medical Stories/Histories): Similar to the previous misclassification, with MEDXS being the second most commonly-annotated top level domain.

- NCOMP (Treatment Noncompliance) was universally misclassified as MEDXS due to semantic similarities discussing medical treatments and MEDXS being overall more popular.
- Finance (FIN), although occasionally very well-classified, could be confounded with Healthcare System (HCS); this error is likely due to the fact that individuals may discuss insurance interaction with trouble finding healthcare services.
- Technology (TECH) and HCS were occasionally misclassified with each other; this phenomenon is likely due to semantic similarity that owes to the technology (e.g., patient portal) being used in a healthcare system-associated setting.

In summary, a majority of commonly-made classification errors were due to untoward semantic similarities between communications that a human annotator could easily distinguish as one top-level domain or the other. The machine learning algorithms all focus on pure lexical semantic content and therefore will be vulnerable to making such errors, especially when one rarer topic is semantically similar to a more common one, causing regression towards the mean/median and thus misclassification as the more common topic.

Finally, also remarkable is (not shown in the above drill-down table) the more impressive recognition of Support (SUP) and Medical Stories/Histories (MEDXS), two non-contextual factor domains detected by all algorithms at $F1 >= 0.600$. The mathematical nature of these algorithms suggests that they tend to classify by regression towards mean or median characteristics, and when certain topics dominate conversation, machine algorithms would therefore regress towards matching communications of uncertain topicality to pre-existing dominant topics. However, as per research documented in other chapters of this work, detection of support-related communications and detection of medical stories and histories (i.e., information directly relevant to the clinical record) were found to be better performed with dictionary-based literal search algorithms, with *F1* exceeding 0.700 in both cases.

The limitations of the research at hand are significant and serve to form a framework on which follow-up research will greatly enhance. The qualitative annotation and thesaurus generation must in the future be re-validated by at least one third-person annotator (preferably one with a clinical background). Similarly, the literature review can be expanded to systematically analyzing non-risk factors that surround AIH patients, although the practical merit and productivity of such a review may be little. Indeed, it is recommended that AIH researchers devote more attention to contextual factors that may not be archetypal disease risk factors; this way, a more complete knowledge provenance of AIH can be formed. Furthermore, the generated thesaurus does not itself constitute an ontology, but the possibility exists for an ontology of not just AIH, but also other chronic rare diseases, can be constructed using the thesaurus as a base to define attributes of affected users.

In thesaurus annotation and machine learning attempts, annotations were performed and communications machine-sorted by fragment and not entire communication. Future research may benefit from (instead of using cross-validation) the utilization of separate testing and training sets. Such an arrangement would allow for the quality assurance analysis of real-world, real-time communication topic detection, as actual users do not split individual communications into fragments by topic.

Many recommendations for future research rely upon algorithm improvement for automated detection as this was the least satisfactory portion of results obtained. The detection reliability was low enough as to preclude any utilization of the algorithms for a corpus-wide classification of contextual factors. Most importantly, proper detection and differentiation of even the top-level domains would benefit by going beyond pure semantic/lexical similarity and will likely require machine learning approaches that take into account other variables, such as context of post vs. comment, metadata of the user, and (for comments), metadata of the user being responded to. More complex algorithms, such as serial ensembles, may also be of use.

In addition, classification by more fine-grained topics (e.g., the 210 topics that rest under the top-level domains) was not attempted here and will require approaches that range from dictionary support to the annotation of thousands of more communications by an expanded research team. Alternately, similarly-annotated materials from other chronic disease support groups could be added to the existing corpus in an attempt to increase the fineness and reliability of automated classification.

The research at hand has demonstrated a valid, thesaurus-based categorization system for classifying the communications (related to contextual factors and otherwise) of users of an AIH-oriented online support group. Contrasts between clinical research opinions and consumer (user) opinions of important contextual factors were successfully delineated by comparing literature review information with annotations of support group user communications.

Machine learning methods were attempted to semantically characterize user-generated content via analysis of word vectors by support vector machine-optimized regression algorithms. While the machine learning attempts did not yield optimal results, they have instead suggested future options to improve classification of communications and therefore enhance detection of contextual factors expressed over online support groups for not just AIH, but also other rare and/or chronic conditions.

CHAPTER 8. CHARACTERIZING SUPPORT GIVEN AND RECEIVED OVER THE AIH GROUP

8.1 Introduction

Patients with rare diseases travel great distances to seek presence of a provider, both for treatment and advice. [4, 5] It is assumed that patients with AIH (and other rare diseases) will inherently face a shortage of support, both socially and informationally and will require forms of support outside the clinic.

I performed a thematic analysis[141] of the literature on support exchanged over health-related social media venues and discovered that there were two dyadic directions (inbound and outbound) [104-107] as well as two types (emotional/social[7, 92, 98, 101, 102] and advice/informational[99, 7, 90, 100]). While dyadic direction of support is a self-explanatory concept, I will specifically define social/emotional support (here abbreviated EMO) as support that does not have objective information or recommendations but instead are essentially kind words given to help someone's emotional status, whereas advice/informational support (here abbreviated AIS) as the exchanging of ostensibly objective advice and information in hopes of support.

I therefore propose to widen the spectrum of AIH patient knowledge gained by this dissertation by analyzing the feasibility of automated detection of types of support provided and requested on AC's AIH-associated Facebook support group and then to characterize the communication structures that surround rare disease social support. In the process, it is desired to estimate the adequacy and appropriateness of support exchanged over the group.

8.2 Methodology

This analysis was performed on the parsed data set that was used throughout this dissertation. Initial classification was performed by a qualitative analysis of communications (posts and replies) on the target Facebook group. From the literature review, the following baseline categories for annotation were derived:

- Offering social/emotional support (OFF.EMO): The provision of support indicated solely for emotional fulfillment and not as practical advice.
- Requesting social/emotional support (REQ.EMO): The requesting of support for emotional fulfillment and not requesting for advice
- Offering advice/informational support (OFF.AIS): The provision of information and/or advice with clear intent to answering a fellow user's question
- Requesting advice/informational support (REQ.AIS): The requesting of advice or information; often phrased as an information-seeking question to the group.

Note: The classification of REQ.EMO, requesting emotional/social support, was theoretically possible but never observed in the annotation set and is therefore not entertained in this analysis.

Other annotation categories, which were derived in a top-down fashion from reading the communications, included:

- Requesting advice/information from AC, the administering colleague (AIAC): Identical to REQ.AIS but directly invokes AC.
- Personal Inquiry (PINQ): It was noticed that in the dyad of support, users would often ask each other personal questions (e.g. *What dose were you on?* or *Are you feeling better?*) in the act of offering support

- Gratefully acknowledging support (GRACK): In reception of support (regardless of type), stating that one is thankful for the support evidently provided. It is assumed that any undirected gratitude expressed in the group is a signal of this category.

Finally, all communications were assessed in a blind secondary run by the first author (AK), re-annotating each communication with the existing thesaurus according to the annotation categories and guidelines above. The intra-rater annotation was considered agreed upon if the annotator agreed with his earlier opinion completely for each given communication. The initial intra-rater reliability was 87.44%. Finally, self-adjudication was performed until an intra-rater agreement of 100% was reached.

A qualitative dictionary was thereby created via annotation of 679 (/900 total) training set communications. The dictionary contains terms that are by the annotator's observation virtually unique to a specific type of support. Some examples of dictionary terms are shown in Table 23.

Table 23. Explored Categories of Support

Category	Example Dictionary Terms
AIAC (Requesting advice from AC)	*[Censored; various terms related to AC's real name]*
GRACK (Gratefully acknowledging support)	*thank, TY, TYSM, thx, thnk, thnx, ty sm, ty vm, tyvm*
OFF.AIS (Offering advice)	*you should, you might, try this, I think you, you need, could work, might work, can help, might help, helped me*
OFF.EMO (Offering social or emotional support)	*hug, prayin, prayer, be ok, be alright, be all right, sorry, hearts, best wish, wish you, sending many, sending much, much love, good thoughts, aww*
PINQ (Personal inquiry)	*is your, do you, does your, what is your, who is your, are your*
REQ.AIS (Requesting advice, not from AC)	*please, does anybody, does anyone, would somebody, can you tell, can you give, can you help, I need help, tell me what*

Note: Some dictionary terms are partial words because the dictionary functioned by substring, and not whole word, matching. A full list of dictionary terms is available in the Appendix.

The algorithm was designed in a fashion where the category with the least observed testing set frequency would be returned first. For example, the dictionary terms in *PINQ* were observed frequently in *REQ.AIS* but the terms in *REQ.AIS* not frequently observed in *PINQ*. Therefore, *REQ.AIS* must be excluded well before *PINQ* will be returned. In addition, the algorithm can only tag each communication with one kind of support, and the priority for tagging follows the priority order in the algorithm. The algorithm is depicted in Figure 13.

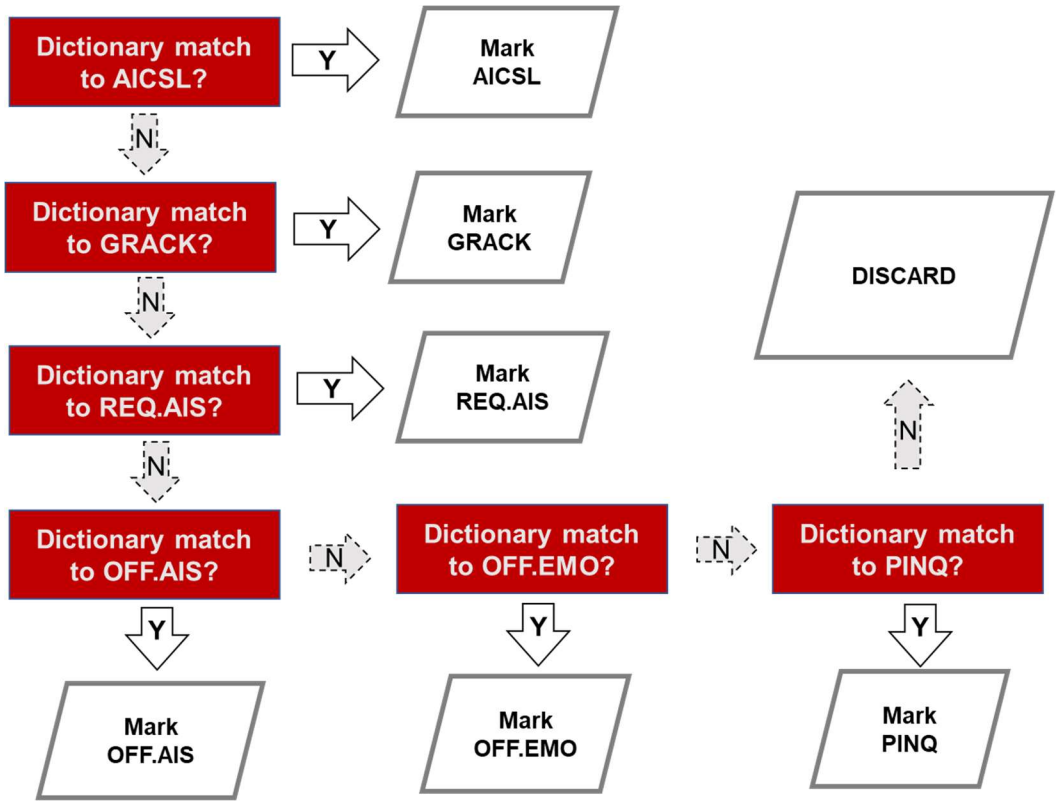

Figure 13. Algorithm Flow for Support Type Detection.[141]

The above algorithm, after validation, was finally performed upon the entire corpus of support group communications in order to characterize its extant types and directions of support.

Finally, in performing an analysis of the support type characterization that can be performed on this corpus, it is decided to correlate in an exploratory fashion network and user-related characteristics with support types evident in communications. This exploratory analysis will assist in validating the classification algorithm and potentially shed new light on support in this rare disease online support group. Data were sorted in order to compare the support-related characteristics of the following: Type of communication (post vs. comment); whether or not the user was deleted from the group at the time of data collection (AC's communications were excluded from this subcount); user role (clinician administrator AC vs. the remainder of the group); user tenure (duration since first communication in the group) *at the time the posting was made* (deleted users excluded).

8.3 Results

676 communications in total were annotated, of which 277 evidenced some form of support. In the second round of annotation, there were 15 instances of support added to communications previously annotated as not involving support. Furthermore, reasons for support were removed from further analysis (although kept on the record for archival purposes). Comparing the training set of 667 communications vs. the testing set of 223, the following reliability measures in Table 24 were observed and are reported as precision, recall, and F1-score.

Table 24. Algorithm Performance: Support Type Detection

Support Type (By Communication)	Precision	Recall	F1 Score	N (GS)
Support vs. not support	0.730	0.692	0.711	78
AIAC (Asking administering colleague, AC, for advice)	0.429	1.000	0.600	3
GRACK (Expressing gratitude for support)	0.533	0.571	0.552	14
OFF.AIS (Offering advice)	0.400	0.250	0.308	24
OFF.EMO (Offering emotional/social support)	0.700	0.700	0.700	20
REQ.AIS (Requesting advice; not from AC)	0.636	0.636	0.636	11

The detection reliability (F1 = 0.711) of support-related communication vs. non-support-related was higher than detection of individual support types. Individually, the best-detected type of support was *OFF.EMO,* the offering of emotional (social) support (F1 = 0.700). *PINQ* (personal inquiry in provision of support) and *REQ.AIS* (requesting of advice and informational support) were acceptably detected with F1-scores of 0.667 and 0.636, respectively. Detection of grateful acknowledgement of support (*GRACK*) was more modest at F1 = 0.552; finally, the detection of *OFF.AIS* (the offering of advice/informational support) was the least satisfactory at F1 = 0.308.

With the reliability of the algorithm assessed, it was then run over entire corpus of communications. Therefore, the corpus-wide occurrence rate of each type of support (and the rate of non-support communications as well) can be, with greater or lesser degrees of error, estimated; these estimates are shown in Table 25. Please note that the double asterisk (**) by

OFF.AIS indicates that relevant data are displayed despite the unreliability of the algorithm in detecting it.

Table 25. Wider Corpus: Support Types Discovered

Support Type	Communications	% of Corpus
AIAC	633	3.32%
GRACK	1,664	8.72%
OFF.AIS**	1,283	6.72%
OFF.EMO	1,186	6.21%
PINQ	462	2.42%
REQ.AIS	1,580	8.28%

In order to better elucidate the social network structures and user properties that underlie support-related communications, support type rates were also broken down via the following criterion sets:

- Communication type (post vs. comment)
- User social network metrics & structure characteristics:

Communications on Facebook™ groups are divided into two structural types: Top-level posts, and comments to these posts. Support varied across communications dependent upon communication type, as shown in Table 26.

Table 26. Types of Support: Posts vs. Comments

		Posts (vs. Comments)				
		OR 5%	**Odds Ratio**	**OR 95%**	**Chi-Square (*df=1*)**	***P*-value**
AIAC	Req. Advice/Info. from AC	5.454	6.445	7.616	613.2	<.001
GRACK	Gratefully Acknowledging	1.565	1.800	2.017	69.4	<.001
OFF.AIS**	Offering Advice/Info.	0.644	0.789	0.968	5.2	.0225
OFF.EMO	Offering Emotional/Social	0.126	0.186	0.276	88.6	<.001
PINQ	Personal Inquiry	0.365	0.540	0.799	9.8	.0017
REQ.AIS	Req. Advice/Info.	4.804	5.408	6.087	937.6	<.001

Posts and comments often had striking differences in the types of support expressed in them. Comments demonstrated a much higher rate of offering emotional support (OR for posts 0.186, 95% CI 0.126-0.276, *P*<.001). Finally, although the finding should be viewed with caution due to low detection reliability; it was noted that the offering of advice and informational support (*OFF.AIS*) was slightly more common in comments (OR for posts 0.789, 95% CI 0.644-0.968, *P*=.0225). Unsurprisingly, these support types are located on the outbound end of the support dyad. The algorithm therefore receives some validation via intuition.

I then analyzed differences in support types by if the user was deleted (left the group, banned from the group, or left Facebook™ entirely, evidenced by a blank entry for user name in the XHTML) at the time of data collection. Note that in this case, all deleted users are processed as one because such users cannot be differentiated. These results are detailed in Table 27, below.

Table 27. Types of Support: Deleted vs. Non-Deleted Users

		Deleted Users (vs. Non-Deleted)				
		OR 5%	**Odds Ratio**	**OR 95%**	**Chi-Square (*df=1*)**	***P*-value**
AIAC	Req. Advice/Info. from AC	2.406	3.223	4.316	68.7	<.001
GRACK	Gratefully Acknowledging	0.599	0.824	1.132	1.4	.2314
OFF.AIS**	Offering Advice/Info.	1.339	1.751	2.289	17.2	<.001
OFF.EMO	Offering Emotional/Social	0.069	0.155	0.348	27.1	<.001
PINQ	Personal Inquiry	0.000	0.000	0.000	14.7	<.001
REQ.AIS	Requesting Advice/Info.	0.933	1.232	1.628	2.2	.1410

Users who were deleted from the network had comparatively high rates of asking AC for his advice (AIAC; OR 3.223, 95% CI 2.406-4.316, *P*<.001). There is also evidence for the increased level offering of advice and informational support (OFF.AIS) in this cohort (OR 1.751, 95% CI 1.339-2.289, *P*<.001), and in contrast, decreased evidence of offering emotional/social support (OFF.EMO; OR 0.155, 95% CI 0.069-0.348, *P*<.001).

Certain observations could be made about user tenure as of the date of communication. There is a drop-off in *REQ.AIS* and *GRACK* combined with the concurrent rise in *OFF.EMO* through increasing tenure (visible in Figure 14). For inbound support types, a U-curve formation is seen (Figure 15), except for AIAC, which shows a very mild increase over increasing tenure. Results are further detailed in Table 28.

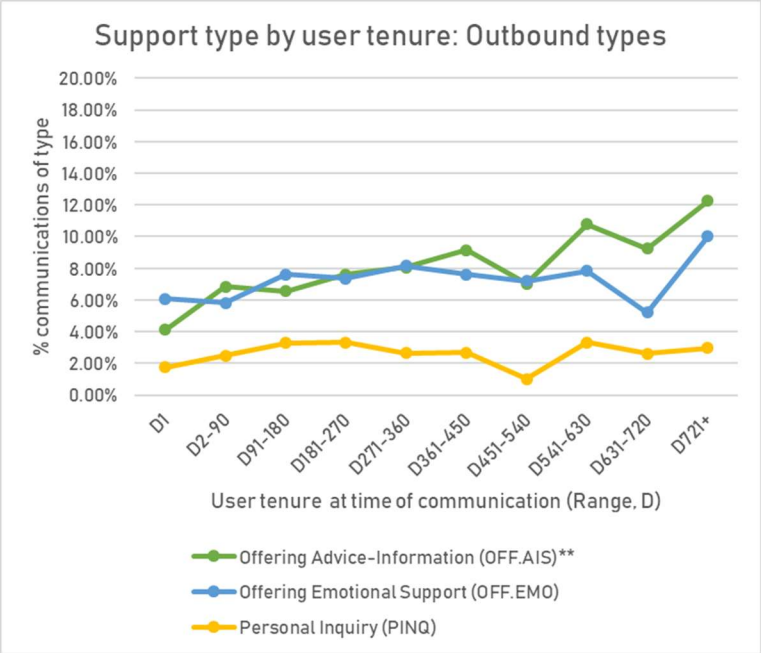

Figure 14. Outbound Support Types vs. User Tenure

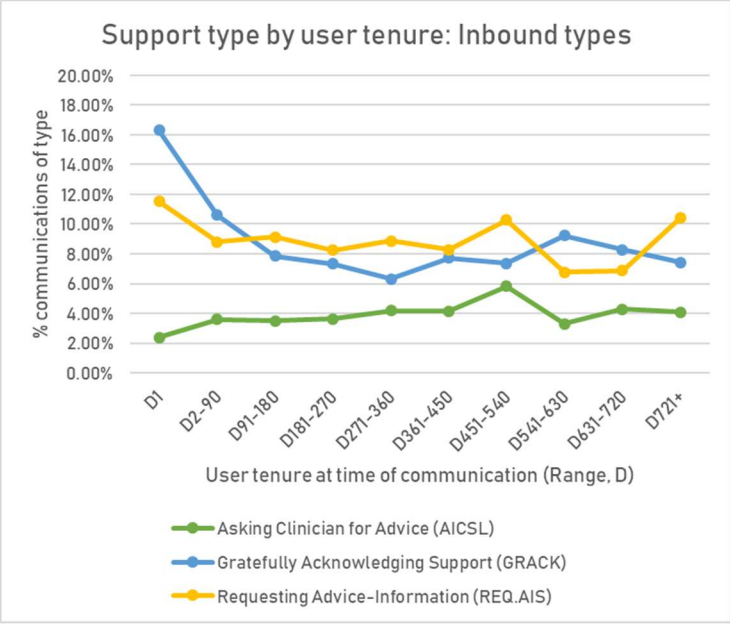

Figure 15. Inbound Support Types vs. User Tenure

Table 28. Support Types by User Tenure (Tabular)

Note: Percentage figures indicate percent of all communications in that user tenure time slot only and do not add up to 100% because non-support communications are excluded.

	User Tenure at Time of Communication % (N)										
	First Day	**D2-90**	**D91-180**	**D181-270**	**D271-360**	**D361-450**	**D451-540**	**D541-630**	**D631-720**	**D721 +**	**Total By Type**
AIAC	2.4% (12)	3.6% (108)	3.5% (83)	3.6% (68)	4.2% (48)	4.1% (43)	5.8% (40)	3.3% (23)	4.3% (22)	4.1% (6)	453
GRACK	16.3% (82)	10.6% (321)	7.8% (186)	7.3% (137)	6.3% (72)	7.7% (81)	7.4% (51)	9.2% (63)	8.3% (42)	7.4% (12)	1,047
OFF. AIS	4.1% (21)	6.8% (207)	6.6% (156)	7.6% (142)	8.1% (92)	9.2% (96)	7.0% (49)	10.8% (74)	9.2% (47)	12.3% (19)	903
OFF. EMO	6.1% (31)	5.8% (176)	7.6% (181)	7.3% (137)	8.2% (93)	7.6% (80)	7.2% (50)	7.8% (54)	5.2% (27)	10.0% (16)	845
PINQ	1.7% (9)	2.5% (76)	3.3% (78)	3.3% (62)	2.7% (30)	2.7% (28)	1.0% (7)	3.3% (23)	2.6% (13)	3.0% (5)	331
REQ. AIS	11.5% (58)	8.8% (267)	9.1% (217)	8.2% (154)	8.9% (101)	8.3% (87)	10.3% (71)	6.8% (47)	6.9% (35)	10.4% (16)	1,053

The final important classification of users is the dichotomy between AC and the remainder of the users. As expected, the offering of advice and informational support was significantly higher (OR 2.824, 95% CI 2.255-3.538, *P*<.001), although this finding must be viewed with caution due to that algorithm's relatively poor performance). AC also appeared have less of a tendency to inquire users *(PINQ)* about their situations (OR 0.401, 95% CI 0.178-0.900, *P*=.0017). In addition, being grateful for support *(GRACK)* was detected in significantly higher amounts (OR 1.477, 95% CI 1.150-1.896, *P*=.0021), in AC's communications compared to the remainder of the user corpus. These results are elaborated in Table 29.

Table 29. Support Types, AC's Engagement.[141]

		AC (vs. not AC)				
		OR 5%	**Odds Ratio**	**OR 95%**	$X_{1,1}$	***P*-value**
AIAC	Req. Advice/Info. from AC	0.192	0.388	0.782	7.539	.0060
GRACK	Gratefully Acknowledging	1.150	1.477	1.896	9.440	.0021
OFF.AIS**	Offering Advice/Info.	2.255	2.824	3.538	88.817	<.001
OFF.EMO	Offering Emotional/Social	0.569	0.820	1.184	1.125	.2888
PINQ	Personal Inquiry	0.178	0.401	0.900	5.251	.0219
REQ.AIS	Requesting Advice/Info.	0.869	1.150	1.522	0.958	.3277

8.4 Discussion & Conclusions

Algorithm performance varied depending on the type of support being detected despite being satisfactory overall. A qualitative *post-hoc* error analysis was performed, comparing annotated posts to the algorithm's results, in order to note weak points in the algorithm.

The poor detection of *OFF.AIS* may be explained by the fact that an annotator can find out when information is being shared in provision of support by the nature of the information while the machine sees no hints that support is being sought. Likewise, the grateful acknowledgement *(GRACK)* of support searches for generic terms such as *thanks* and its synonyms, creating the observed probability of misclassifying a user who was thankful, but for something other than support.

In the receiving end of the support dyad, the most commonly observed types of support were the requesting of advice/information (REQ.AIS) and being grateful for support (SUP.GRACK). It is more significant that requesting advice and informational support is very common; such an observation strengthens the potential utility of detecting user requests for

advice on an active personal health platform with the goal of delivering the user customized patient health education materials. Furthermore, the requesting of advice in this group corresponds with literature evidence that this type of support is a major component of online health-related support groups, with multiple sources[92, 98, 100] discussing the importance both of giving and receiving this type of support.

Also noted was the offering of emotional/social support (OFF.EMO), which is a type that is likely used in diseases (including AIH) that have no definitive cure. Literature has observed this type of support being used commonly by patients of pulmonary fibrosis (PF)[92] and the incurable status of breast cancer survivorship[103]. Less common was *PINQ* (personal inquiry in the provision of support), indicating some weakness in the support dyads formed. Deliberately asking the administering colleague (AC) for support *(AIAC*) was also less common than general requests for advice, bucking the trend of the clinician's assumed dominance in the group.

Posts and comments differed significantly in types of support discovered; comments were biased towards the outbound side of the support dyad. In contrast, the other three types of support, coincidentally located on the inbound end of the support dyad, are noted at a higher rate in posts, creating further intuitive validation of the algorithm. The algorithm therefore receives some validation via intuition.

The more inherent characteristics (not related to social network structures) observed were user deletion from the network (due to leaving the group, leaving Facebook entirely, or being banned from the group due to conduct violation), tenure of the user as of the date of a communication, and the role of the user (administering colleague AC vs. all other users). Unless otherwise stated, the conclusions made in this section are speculative and require further information as specified.

Users who were deleted from the network more often asked AC for advice. It is possible but not confirmed that they may have lost interest in the group after asking a single question and receiving AC's advice. If the conclusion of increased *OFF.AIS* discussion is validated by

algorithm improvement, the increased level of communication in that domain in this cohort may indicate the presence of fraudulent information (offered as advice) for which the user was banned; however, further investigation into the *OFF.AIS* posts is indicated in order to gauge user attempts at providing fraudulent and possibly harmful information as advice.

The tenure of the user at time of communication (i.e., the number of days they have been in the group at the time they made the made the communication) is another facet of study due to the need to explore changes in support involvement across time. This feature implies not just trends about communications, but also those inherent in users themselves. The drop-off in *REQ.AIS* and *GRACK* combined with the concurrent rise in *OFF.EMO* through increasing tenure (visible in Figure 3) may be intuitively explained by the fact that users tend to shift away from requesting and receiving support and instead towards giving others support. The potentially paradoxical tail end of the U-curve seen for the inbound support types (Figure 3) may be explained by long-time users facing more severe disease and having more support needs; this is a phenomenon that requires further investigation. Another paradoxical increase in one type of inbound support (requesting from AC, *AIAC*) can only be speculated by the idea that users become more comfortable asking AC for advice the longer they stay in the group.

AC was significantly more likely to offer advice (OFF.AIS), with relevant limitations for the algorithm. More unusual is the fact that he appeared to have been more prone to offering gratitude (GRACK), although this finding may have been due to him thanking users for their participation and not for their advice or emotional support. Finally, judging by the lack of AC's PINQ content, it appears that AC might be offering information without becoming too personal in investigating individual patient histories, suggesting that his conduct and advice are appropriate and legal in nature.

The discovered limitations of automated support type detection form a significant framework upon which future research can be based. Detection of an important type of support, the offering of advice and information *(OFF.AIS*) was inferior to detection of other support types. It is possible that future research can investigate communication structures in order to increase detection of *OFF.AIS:* For example, health-related information shared in reply to a post that requested advice should also be (but is not) tagged as *OFF.AIS.* Similarly, detection of grateful acknowledgement of support (*GRACK*) could be enhanced by searching for previous outbound support within the post's comments.

Due to reasons of scope, the detection and categorization of support utilized also did not explicitly classify the *reason* support was being given or solicited. For example, it is not known if emotional/social support was given due to another's grief or if informational support was given to a question regarding medication side effects. Further dictionary-based classification of communications (for REQ.AIS) and of neighboring communications (for offering subtypes) may elucidate reasons further.

In addition, while detection of requesting advice and informational support (*REQ.AIS)* was sufficient, Facebook does not allow an automated intervention in response to detection of this type of support need. It is therefore recommended that this algorithm in the future be implemented in an active patient health platform where users can post and comment with their communications (consentingly) monitored and then relevant advice in the form of patient education materials returned to them based on the remainder of their communication content.

Finally, there are portions of the dictionary that are customized to the domain of AIH and even this group in particular. Specifically, the synonym set for detecting invocation of the clinician (*AIAC*) functions based upon AC's actual name, and future detection of clinician invocation should be customized to the clinician(s) in the respective group. In addition, there is no dictionary for the detection of emotional support requests (*REQ.EMO*) because the support type was never observed in the training set.

The research at hand has demonstrated a pilot, proof-of-concept methodology for detecting certain types of support exchanged over an online support group for the rare disease autoimmune hepatitis (AIH). The algorithm, which functions by a dictionary that is enforced by a binary tree, is able to detect with satisfactory reliability the requesting of advice, the invocation of the administering colleague, personal inquiry in the provision of support, and the offering of emotional (social) support. Detection of grateful acknowledgement of support is modest, and the least detectable support type was the offering of advice and informational support.

Through this methodology, the dynamics of support across cohorts of users and time has been successfully described, with both intuitive as well as surprising findings generated. The generated algorithm finally has the potential to enhance a patient health platform that can be attentive of when a patient is seeking information (requesting advice/informational support) and return via further dictionary searching appropriate patient educational materials.

CHAPTER 9. CONCLUDING SYNTHESES & FINAL DISCUSSION

9.1 Synthesis of Methods & Findings

Across all aims and chapters comprising this dissertation, I have retrospectively identified the following themes pertinent to methodology: Qualitative Classification, Methodology, Computational Methodology, Reliability, and Reusability of Method. The following synthesis brings together these themes across all research aims and chapters in order to form a suitable conclusion for the research work at hand.

The first theme across which the aims will be compared is qualitative classification methodology, which for purposes of this discussion consist of the subjective criteria and objective level by which communications are classified.

Classifications of information exchanged and support, despite their temporal distance in when research was performed, both utilized a thesaurus-based methodology in classifying the subjectivity of user communications. The research in both aims was initiated from a *top-down* read of user communications, which was followed by internal validation. The research in classifying types of information shared was more consistent with the initial research in thesaurus classification of patient contextual factors[67]; later on, I renamed *Economic* to *Financial (FIN)* and *Health Behaviors* to *Self-Treatment Stories (STX).* More significantly, the original *Psychological* domain was largely re-classified into the psychological sub-domain of *Quality of Life Factors (QOLS.PSY)* due to results from the literature review in the latter aim.

Somewhat similarly, issues of signs, symptoms, and comorbidities – including pain, injury, and laboratory test results were mapped not just to distinct SNOMED-encoded clinical entities, but to higher level coarser SNOMED entities that effectively clustered the finer SNOMED codes, in effect forming a hierarchy similar to that found in a thesaurus. In contrast, xenobiotic/drug usage classification was done with extremely fine granularity (i.e., to compound

identity and not therapeutic class), chiefly in the interest of potential hepatic effects of these chemicals. One chief exception is that classification of xenobiotic products as AIH-related therapeutics was also informally performed in order to qualitatively validate the study's results.

In contrast to qualitative methodology, which in this dissertation is human-curated, computational methodology is automated. Computational (defined here as automated computing-based) methodology differed significantly across aims, with clusters of aims having similar methodology utilized. Research methods used in overall group characterization and contextual factor analysis were similar in that they both used *black box* computational approaches, allowing *off-the-shelf* machine learning software products to analyze the data with relatively little researcher interference. In the former, the lack of existing knowledge of patient communication content drove the choice of methodology; in the latter, it was the known sheer complexity of data (with 172 variables found qualitatively).

In contrast, detection of drug (xenobiotic) usage, signs/symptoms/comorbidities, and types of supports exchanged all utilized recognitions of named entities (named-entity recognition, aka dictionary-based search or simply NER) combined with rule sets. A fixed and typically lower number of observed categories allowed for qualitative synthesis of customized dictionaries that could then be run across the wider corpus of communications. The exception to this logic was the analysis of AIH patient xenobiotic (drug) intake, chiefly due to the fact that it was intended to detect the consumption of over 7,500 xenobiotic products (and in fact detected consumption of over 250). Because of the US Food & Drug Administration (FDA)'s Orange Book[131] of generic and brand names – and an added step to account for misspellings – an NER-based search was still feasible.

Computational methods, in turn, are judged by reliability. Reliability, for purposes of this discussion, is defined as the harmonic mean (*F1*-score) of precision and recall for any detection algorithm. Reliability varied significantly across aims and aspects sought in each aim. Literature thresholds for *F1*-based reliability vary greatly; for purposes of this discussion,

moderate reliability will be considered 0.500 <= *F1* <= 0.670 and high reliability *F1* > 0.670. Below is a reference table of *F1* reliability values across aims and aim components. A reliability summary is demonstrated in Table 30.

Table 30. Algorithm Performance Summary: F1-Scores

Aim	Component	*F1*-Score	Notes
B1	All drug intakes	0.731	Correct xenobiotic & usage by user
B2	Pain & Injury	0.667	Correct injury & location by user
	Signs, Symptoms, & Comorbidities	0.729	Correct sign/symptom/comorbidity class by user
	All Lab Test Results	0.639	Correct test & result by user
B3*	All Contextual Factors	0.317	Correct factor by communication; best algorithm
	Finance (FIN)	0.529	By communication; best algorithm
	Social (SOC)	0.545	By communication; best algorithm
C*	All	0.730	By communication; is support vs. not support
	Offering emotional support (OFF.EMO)	0.700	By communication
	Personal Inquiry (PINQ)	0.667	By communication
	Requesting Advice (REQ.AIS)	0.636	By communication
	Asking AC for support (AIAC)	0.600	By communication
	Expressing gratitude for support (GRACK)	0.552	By communication
*Not all data are shown for B3 or C due to the high number of variable types discoverable			

Some user facets evidenced high reliability. Drug/xenobiotic usage, down to the correct compound (B1), was reliably detectable, as were an array of signs, symptoms, and comorbidities (B2). The chief similarity between these facets (and analyses) is that they were calculated by user and not by communication. The calculation, although appropriate for the purpose of the respective study, may implicitly deflate the reliability achieved in per-communication analyses. However, some communication-specific metrics evidenced high reliability; such metrics included the detection of communications that were vs. were not support-related, and also the offering of emotional support (OFF.EMO). All of these facets were explored by named entity (NER;

dictionary-based) search, indicating better performance of this algorithm vs. *black box* machine learning approaches.

More modest reliability, in contrast, was found for pain and injury (B2), all lab test results (B2), finance (FIN) as a contextual factor (B3), social history as a contextual factor (B3), and several types of support (C). Except for those in B2, these were all sought on the communication level, potentially deflating the reliability estimates. Many of these facets were sought using a *black box* machine learning approach, and it is important to consider that such an approach, while often the only option, may not yield the results that an NER/dictionary-based one would.

Reliable algorithms are useful, but only if reusable (i.e., *interoperable* in separate but related use cases). The reusability of methods varied by the exact algorithm and aim. In this sub-section, the reusability of sufficiently-performing ($F1 >= 0.500$ in all cases) algorithms is elaborated.

Some utilized algorithms are likely to be highly reusable because they operate on assumptions that are common to all social media communication content. Latent Dirichlet Allocation (LDA) operated via the Machine Learning Language Toolkit (MALLET)[119], being an *off-the-shelf* and *black box* solution, is by nature reusable (and not a product of this research). More significantly, however, the named entity-rules-based hybrid search for xenobiotic/drug usage, which is original to the research at hand, would be usable in other corpora of user-generated online support group content. As such, it may be useful for *pharmacovigilance* analyses in commonly-available social media information. Finally, detection of certain forms of support (in particular the requesting of advice and information, REQ.AIS), was sufficiently reliable and the algorithm dictionary used for it could potentially be implemented in real-time patient health platforms to help determine when a user was actively seeking health advice.

The algorithm for detection of contextual factors, in its better performing top-level domains, would require some modification in order to be reusable across different disease

domains. Normalization takes into account only AIH treatments and would have to be modified to normalize common treatments for the disease at hand.

Finally, the least reusable methodology is likely that which detects signs, symptoms, and comorbidities. While a few signs and symptoms (e.g., mood issues) will be found in corpora across many disease domains, diagnosed comorbidities in the dictionary are almost exclusively autoimmune in nature. The dictionaries for this algorithm would therefore require that extensive re-annotation of any new corpora be performed.

9.2 Synthesis of Other Limitations & Future Research

The purpose of this sub-section is to synthesize, across the different aims, common limitations *that were not covered in the synthesis of findings* and extend these limitations to what future research will best help this research cause. These limitations include those pertaining to semantic ambiguities, user underreporting, small sample sizes, and the lack of an associative/predictive model.

The first general limitation to be discussed is semantic ambiguity, which for purposes of this discussion is defined as the literal lexical content of a communication deviating from its logical content. Semantic ambiguity most often caused problems in *black box machine learning* approaches. For example, *The aza [azathioprine] keeps me in bed* was annotated as *QOLS.NEG.NRG* (poor energy as a quality of life factor) but because of the presence of a drug name was matched by the machine learning algorithm to *MEDXS* (discussion of medicine intake and medical symptoms). Such a deviation may also occur when a user is commenting on another user's post, and either an annotator who understands disease context – or the other user's posting text – is required to determine the actual logical meaning of a communication.

Other portions of this research discovered semantic ambiguities, albeit at a much lower rate. Rare cases of drug US brand names were within 75% spelling similarity to commonly used

words. Further discussion on this type of error is available in Chapter 5's Discussion section, to which the reader is encouraged to refer.

Future research to mitigate semantic ambiguity as it results from *black box* algorithms will require enhancing the data given to the machine. A clear potential exists to enhance comment communication TF-IDF data with the TF-IDF data from respective parent posts. In addition, the identity of the user who was replied to may also assist machine learning-based classification of comments. Previous mistakes can also be coded and given to the algorithm. In the previous example, another field can supplement the communication's TF-IDF matrix; this field can state whether the machine earlier classified it correctly, and if incorrectly, state the incorrect class.

Even if lexical content mapped directly to the semantic, there still will exist *user underreporting,* which for purposes of this discussion is defined as the online support group users not reporting every problem experienced, every question thought of, every answer considered, and every medication taken.

Measures of underreport could be quantitatively defined with regards to the detection studies for drug (xenobiotic) intake and signs/symptoms/comorbidities (SSCs). For drug intake, it was noted that 337/1052 (32.0% of) users declared intake of azathioprine. In reality, a majority of AIH patients (50.0% or greater) are treated with this drug[41, 121, 142], implying an underreport rate of at least 36% for azathioprine intake. Underreport of intake of non-AIH-associated medication is also a likely scenario, with (for example) the very common pain reliever Ibuprofen (ADVIL, MOTRIN, etc.) only reported by 2.0% of group users as taken.

In the studies to detect SSCs, the underreport of symptoms and comorbidities, including cirrhosis, was hypothesized. Conservative estimates have put the prevalence of cirrhosis *at AIH presentation* of 25%[31] with rates in established AIH much higher. In contrast, it was found that only 10.6% of online support group members declared having cirrhosis. The gap observed for

comorbidities less associated with AIH was even more striking: For example, 222/1052 (21.1% of) users reported having had an infectious disease; the actual rate is most likely closer to 100.0%.

In these cases, future research, if restricted to social media, may not enhance detected report rates because the algorithms in question already possess sufficient recall; solutions may exist to mitigate these issues. Assuming that less active users (with lower communication degree) are the chief under-reporters, the data set can be trimmed to either users making over a threshold number of communications or users reporting a minimum threshold number of clinically-related factors (signs, symptoms, comorbidities, and drug intakes). Preliminary *post-hoc* analysis of the data, conducted after the under-reporting was recognized, demonstrated that users who reported at least ten clinically-related factors (N = 151) showed a 51% rate of claimed azathioprine intake and a 39% rate of cirrhosis diagnosis claims.

Furthermore, for signs, symptoms, and comorbidities, it is possible but unproven that xenobiotic/drug usage can be used *in absentia* of their primary disorder of treatment as surrogate markers of pathology (e.g., antidepressant intake used as a surrogate for mood issues; ibuprofen intake used as a surrogate for physical pain). However, the best practice for optimal data coverage for these clinical factors will likely be traditional researcher-patient (or provider-patient) interview sessions and surveys, as evidenced in previous research.[88]

The underreport phenomenon begets an issue of small sample size. Although the sample sizes discovered and utilized were generally superior to those obtainable in real life studies of rare disease patients, they still restricted many of the studies at hand to pilot status.

In particular, the sections on drugs/xenobiotics and SSCs were limited to reliability validation performed on the user level, with only 35 *gold standard* testing set users. The *F1*-reliability scores generated, while often high, are nonetheless subject to standard error. However, it is also noted that these reliability scores were generated across multiple classes (i.e., multiple drugs and multiple signs, symptoms, and comorbidities), technically increasing the sample size between 2- and 10-fold depending on the evaluand being studied.

When the studied communications were studied by communication and not by user (and further broken down into fragments), the sample size was also greatly increased as per the research portions regarding contextual factors and exchanged support. 890 communication fragments were analyzed for contextual factor and communication topic content; in detecting support content, the testing set was in excess of 200 communications. Nonetheless, in the case of rarer top-level contextual factor domains as well as rarer support types, not only did the sample size become small, but the responsible algorithms also tended to perform poorly.

Future research therefore should focus on the annotation of larger portions of the corpus as well as annotation of other, similar chronic disease support group corpora in order to gain sample sizes sufficient enough to narrow the margin of error of *F1*-reliability scoring and potentially improve the performance of machine learning-based algorithms.

Despite the valuable self-reported information gained (often with reasonable sample size and a high degree of reliability) from the work at hand, association of such factors is not entertained in this dissertation. In fact, this limitation was deliberately imposed because it was not known whether all such factors could even be reliably detected over social media support group postings. The only associative analysis performed was on the characterization of support exchanged; correlations were performed between support types and various user facets.

Future research to remedy utilizing the data that was successfully derived through this dissertation's research could potentially be used to construct associative and predictive models for disease. Clinical factors such as xenobiotic/drug intake and signs/symptoms/comorbidities (including pain and lab test results) can all be cross-associated to determine hundreds of pairwise and multiple-logistic association strengths from the data.[43] Contextual factors, on the other hand, would likely require better detection algorithms in order to be featured in associative analysis.

9.3 Final Conclusion

The research at hand has demonstrated satisfactory-to-high reliability in detecting general types of health information shared; xenobiotic/drug intake; experienced pain, signs, symptoms, diagnosed comorbidities, and lab test results; types of support exchanged; and select patient contextual factors, all over an online autoimmune hepatitis (AIH) support group.

It is expected that this research has generated significant impact by proving pilot feasibility of extracting clinical and non-clinical patient factors from social media. In addition, the research has successfully surveyed an AIH patient/caregiver "digital cohort"[85] (p. 618) of previously unattainable size (1,052 users, of which 687 mentioned at least one clinical factor). The research has also impacted knowledge of non-clinical factors experienced by support group users; in particular, exchange of support was well-characterized and pilot methodology for detection of an array of patient contextual factors was established.

Most importantly, the research at hand was necessary to facilitate future research in rare and chronic diseases. Many of the developed algorithms are expected to be cross-domain in nature, executable on any support-oriented social media communication corpus. The data derived from the more clinical research portions (e.g. xenobiotic/drug intake; symptoms) can be used in associative models to shed more light on potential drug-disease correlates, putative drug adverse events, and symptom-symptom correlations. Finally, the attempted detection of contextual factors creates a framework on which its own performance can be improved and used in wider disease corpora.

APPENDIX I. TOPIC MODELLING-GENERATED TOPICS, WITH WORD BASKETS AND AC'S CLASSIFICATIONS

Used with permission from research of the author's original work: Kulanthaivel A, Lammert CS, Jones JF. (2018). A novel approach using social media to investigate patient-centric data in autoimmune hepatitis. (Poster). Washington, DC, USA: *Digestive Diseases Week 2018*.[116]

Mallet ID	Strength	Topic Keywords	Clinician (AC) Group Classification
0	0.28%	euro acirc checked brvbar rosacea eye supplements set link gel hole pressure draw intolerant herbs metals janet hair broken bleeding	Alternative treatments
1	0.66%	pred post teeth yeah lol yep kids dentist fine bad weeks stuff migraines allowed pre refused oral gums house chest	Treatment side effects
2	0.24%	chirrohis predisone mine lol teeth stage hepatologist yrs drs imuran alot agree heard fine alittle vertigo lenox fatty awesome idk	Treatment side effects
3	0.74%	pregnancy pregnant baby risk flare pregnancies diseases doctor healthy gum postpartum qualify related born boys income fluid adhd birth ginger	Pregnancy
4	0.70%	quot inflammation fibrosis low post point levels therapy igg increased impact evidence speaking increase small specifically immunosuppression activity speak talk	Treatment
5	4.54%	doctor doctors liver taking blood diagnosed prednisone medication medications skin check good make numbers insurance diseases dont caused prescribed lab	Treatment
6	2.16%	aih liver disease patients study risk treatment patient follow pbc diseases medication things history based drug members clinical community medications	Treatment
7	4.20%	liver biopsy normal test high results negative inflammation damage blood elevated positive enzymes tests tested diagnosis interesting showed upper treatment	Treatment goals
8	4.50%	great group information support autoimmune important questions hepatitis sharing make share rare research care long interesting treatment glad agree news	Support groups

Mallet ID	Strength	Topic Keywords	Clinician (AC) Group Classification
9	0.89%	emoticon smile frown feel calcium thistle coffee magnesium wink vit pains hormone thankful vomit hormones supplements shakes request mood milk	Alternative treatments
10	0.85%	transplant spleen disease platelets low enlarged count blood wasnt aza diabetes case sugar itp removed metformin due fluid weight pred	Comorbid conditions
11	0.71%	cellcept prednisone switched fibromyhellgia works imuran stomach raise mgs easier rituximab sucks teaching hours adjust cyclosporin shots diarrhea fantastic copd	Treatment
12	1.24%	immune system people trigger med heard illness specific stress steroid track steroids heres term amount general bone developed doses link	Pathogenesis
13	0.76%	yuml eth tilde rsquo trade acirc oelig euro bull lsquo permil iuml amazing cedil news rsaquo bdquo curren raquo cent	*Entirely stop characters; discarded*
14	3.47%	aih lammert craig hope hepatologist side diagnosis time flare effects lfts research feeling experience opinion dont antibiotics conference question agree	Research
15	7.30%	years aih ago side time months aza effects drug life times dose put live week hair thoughts age past drugs	Treatment side effects
16	0.10%	yrs count pancreatitis stage white azathioprine hashimotoshypothyroidism weekly nite ulcers landed shots drs neupogin reply youssef trace switzerland burning daughter	Treatment side effects
17	4.29%	prednisone imuran doc started taking month labs numbers flare start test levels lfts docs fatigue results week weaned hoping stopped	Treatment
18	0.82%	son type immune auto year diabetes diseases psc sons dylan headaches chronic hes gastritis tacrolimus doesnt overlap prednisolone melatonin couple	Comorbid conditions
19	5.86%	diagnosed aih week weeks ive azathioprine months hospital bit daily levels question great wondering due yesterday curious appointment type doctors	New diagnosis
20	0.32%	mayo provider lobe medical necrosis left local studies scan lab live patient finland mass clinic story phos cardiac leg dismiss	Comorbid conditions
21	0.30%	products daily results immunoglobulin copies level tests globulin hepatologist herbal nurse mgm quit benadryl regular hurt yrs medical lotion sarna	Treatment
22	11.20%	dont good pain people find feel disease read family found hard lot love understand told makes havent luck fatigue helps	Caregivers

Mallet ID	Strength	Topic Keywords	Clinician (AC) Group Classification
23	1.32%	weight brain lost lose fog gained lbs pounds easy illness fun taste caused matter sucks hungry loose carb fat sense	Disease associated phenomenon
24	13.51%	liver back meds year time work day days bad mine feel didnt blood started issues ago weeks things long night	Treatment side effects
25	1.47%	god hope prayers praying pray happy blessed listen itching bless wonderful thankful skin awesome healing merry amen life finally worry	Religion
26	0.49%	amp medical fibroscan dna awareness stress yrs drs area treatments community denial leg advil reducing protein educate rub monday reaction	Research
27	0.33%	levels mouth thyroid dry possibly bump inr egd generally sodium pretty related hang yeah mmp occur ranges learned people facility	Comorbid conditions
28	1.64%	alt normal ast range flare budesonide numbers labs remission immune lab slightly lower doesnt routine completely flares gastro period short	Treatment goals
29	0.33%	budesonide lammert steroids myfortic probiotics medicine mouth issue dry cleveland lucky equine puppy depression insurance acupuncture moms nodes ursodiol missing	Alternative treatments
30	0.36%	lupus plaquenil rheumatologist symptoms sjogrens arthritis rheumy psoriatic inflammatory add joint mos suppressant azathioprine debbie syndrome hands insulin ais symptom	Comorbid conditions
31	0.70%	daughter daughters shes school prograf pred diagnosed takes tacrolimus drs advagraf vitamin college day worried child childrens reduce mum ruptured	Treatment - pediatric
32	0.64%	insurance drugs enzymes pharmacy health store gastroenterologist price pay problems change yoga subject paying cost grocery deductible complete estrogen california	Medication payment
33	0.38%	conference aiha weekend meeting indianapolis laurie indiana hellip hope hotel email june auction ebv attending indy relapse wonderful aware considered	Support groups
34	0.21%	lfts study trial dxd posted part florida love update feeling women joy acirc anxious ill updates asked injection ladies spif	Research trials
35	1.30%	water drink tea helps green yrs drugs dont iron bladder eat cool gall ice cancer switched wont coconut sensitive salt	Alternative treatments
36	0.13%	cyclosporine dosage empathy ddd meditation addition turmeric sympathy myth cravings spine condition alcohol spec eye handle fry compassion physiologically toledo	Alternative treatments

Mallet ID	Strength	Topic Keywords	Clinician (AC) Group Classification
37	1.19%	ive youre ill wow lucky youve hey nurse theyre increased wondering relief lot rhuematoid yep beginning imagine hes shouldnt breathing	Comorbid conditions
38	2.89%	hep aza pred months dose remission numbers taking day higher tests gain treatment doesnt concerned added weight small raised aware	Treatment goals
39	0.65%	hashimotos hip knee replacement legs lucky surgeon shoulder synthroid spine honest shots replaced permanent convinced occasional tongue artist videos knees	Comorbid conditions
40	0.75%	pbc ursodiol urso overlap aihpbc stage bile primary hepatologist clinic mayo hepa itching biliary ama skin intense ducts referred bloodwork	Comorbid conditions
41	0.40%	tacrolimus generic medicare insurance medical approved prograf denied med pay fda seattle insulin drug send city pocket virginia vaccine supply	Treatment
42	1.36%	diet eat free gluten food sugar foods oil eating vitamin paleo avoid medicine supplements dairy clean processed aip alternative organic	Alternative treatments
43	1.11%	cirrhosis liver stage transplant portal severe varices ultrasound procedure meld hypertension hepatic failure bleeding considered lactulose encephalopathy standard ammonia diabetes	Disease side effects
44	0.10%	patients question good data aiha number tests approach important commonly time related conference media approaches association typically findings ldquo	Research
45	2.41%	pain joint diagnosed fatigue muscle disability fibromyalgia extreme nausea feels due joints arthritis vomiting october experience gallbladder failure level healing	Comorbid conditions
46	1.43%	quot early switched removed heard doc appointment youre stronger touch wear cup alternative agree connective spot lucky cataracts put tub	Treatment side effects
47	0.58%	bloods prednisolone consultant coeliac bit scan craig scans understand steroids biopsies sbquo alt tablets london reduced perspectum sjorgrens heal mri	Research
48	2.94%	autoimmune disease diagnosis symptoms hepatitis common months treated hepatologist helpful diagnosed azathioprine plan issues conditions chronic easy daily genetic list	Treatment
49	5.30%	told sick husband couple didnt hospital thing flu ill shot asked lol pretty times good gave drink eat home give	Caregivers

APPENDIX II. LIST OF ALL DICTIONARY TERMS CONFIRMING DRUG INTAKE.

Note: Underscores indicate spaces in real life communications.

Negations

"never_took", "never_had", "nt_take", "cannot_take", "never_tried", "nt_have", "not_given", "not_prescribe", "not_been", "not_prescribing", "refuse", "never_taken", "you", "your", "you_re", "youre", "not_prescribe"

Caregiver Indicators

"he_","daughter", "son", "father", "hubby", "wife", "hes", "my_daughter", "my_son", "my_hubby", "my_wife", "my_child", "year_old", "toddler", "boy", "girl", "baby"

Caregiver Present Intake

"takes", "is_taking", "takes", "is_on", "trying", "on", "being_given", "was_prescribed", "was_given", "hates", "can_take", "chooses", "uses","upped_his", "upped_her", "hates"

Caregiver Passive

"gave_him", "gave_her", "gave_my", "prescribed_him", "prescribed_her", "gave_her", "prescribing_him", "prescribing_her", "prescribing_my", "giving_my","giving_her", "giving_him", "gives_him", "gives_her", "gives_my", "started_him", "started_her", "raised_his", "raised_her", "had_her", "had_him", "prescribe_her", "prescribe_him"

Patient Present Intake; Patient Suffixes

"_on", "anyone_else", "any_one_else", "trying", "im", "i_m", "i_am", "im_on", "using", "taking", "take", "started", "been_on", "was_given", "has_me", "makes_me", "makes_my", "me_on", "up_my", "could_take", "missed","titrate", "taper", "lower", "raise", "raised", "lowered", "tapered", "titrating", "lowering", "tapering", "raising", "causes_me", "causes_my", "causing_me", "causing_my", "just", "hate", "hating", "prescribing", "gives_me", "prescribe", "dont_like", "do_not_like", "don_t_like", "added","weaning", "choose", "have", "_can", "starting", "going_on", "use", "using", "now", "not_to_mention", "being", "starting", "stopping", "will_start", "will_stop", "prescribed", "prescribing", "prescribes", "can_only", "can_do", "have_me_on", "still_on", "allergic_to"

Patient Past Intake

```
"took", "was_on", "stopped", "tried", "off", "was_on_", "d_been_on",
"hated", "gave_me", "had_me", "nt_tolerate", "not_tolerate", "gave_me",
"no_more", "never_again", "ve_taken", "made_me", "made_my",
"caused_me", "caused_my",  "did_not_like", "didnt_like", "did_nt_like",
"d_taken", "was_taking", "had_been", "chose", "tried", "used",
"was_put", "took", "no_more", "used" "had_me_on", "discontinued",
"allergic_to"
```

APPENDIX III. LIST OF XENOBIOTICS/DRUGS CONSUMED WITH FREQUENCIES

Note: Only xenobiotics/drugs with a detected rate of N >=5 are shown.

	Xenobiotic	N
#	6-Mercaptopurine	48
A	Acetaminophen	43
	Allopurinol	6
	Alpha lipoic acid	7
	Alprazolam	5
	Amitriptyline	6
	Amoxicillin	20
	Aspirin	10
	Azathioprine	338
	Azithromycin	6
B	Beta blocker NOS	6
	Budesonide	88
	Bupropion	11

	Xenobiotic	N
C	Calcium	31
	Chloroquine	5
	Clavulanic Acid	14
	Clindamycin	12
	Cortisone	8
	Cyclosporine	16
D	Diclofenac	5
	Diphenhydramine	15
	Diuretic NOS	27
	Doxycycline	8
	Duloxetine	7

	Xenobiotic	N
F	Fluoxetine	7
G	Gabapentin	14

	Xenobiotic	N
H	Hydroxychloroquine	17
I	Ibuprofen	20
L	Lactulose	16
	Levonorgestrel	6

	Xenobiotic	N
M	Magnesium	23
	Melatonin	11
	Metformin	9
	Minocycline	7
	Morphine	8
	Multivitamin	8
	Mycophenolate	99
N	Naproxen	9
	Neomycin	6
	Nitrofurantoin	5
O	Oxycodone	8

	Xenobiotic	N
P	Prednisone Prednisolone	284
R	Rifaximin	6
S	Sertraline	5

	Xenobiotic	N
T	Tacrolimus	36
	Tramadol	11

	Xenobiotic	N
U	Ursodiol	58
V	Vancomycin	5
	Vitamin C	8
	Vitamin D	42
Z	Zolpidem	6

APPENDIX IV. LIST OF ALL SNOMED CODES WITH PARENT TERMS

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Abdominal Pain	21522001	Abdominal Pain	21522001	
LUQ Pain	301715003			
RUQ Pain	301717006			
Abnormal Hair Growth	85305001	Abnormal Hair Growth	85305001	
Alopecia	56317004			
Alopecia Universalis	86166000			
Abnormal LFT	707724006	Abnormal LFT	707724006	*In the setting of AIH, an abnormal LFT is always assumed to be high.*
AIH Flare	408335007.FLARE			
Bilirubin Elevated	14783006			
Elevated Alkaline Phosphatase	274770006			
Elevated ALT	707724006.ALT			
Elevated AST	160931000119108			
Elevated Bilirubin	14783006			
Elevated CPK	432352001			
Elevated GGT	707724006.GGT			
Elevated LFT	707724006			
Unstable ALT	75183008			
Abnormal Menstruation	386804004	Abnormal Menstruation	386804004	
Amenorrhea	14302001			
Menopause	161712005			
Premature Menopause	373717006			
Abnormal Stool	271840007	Abnormal Stool	271840007	
Acholic Stool	70936004			
Diarrhea	62315008			
Acne	11381005	Acne	11381005	
Steroid Acne	201222006			
Advanced Stage AIH	408335007.ADV	Advanced Stage AIH	408335007. ADV	
Stage 3 AIH	408335007. STAGE3			
Stage 4 AIH	408335007. STAGE4			
Excipient Allergy NOS	438784000	Allergy Disposition	609328004	

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Allergy Disposition	609328004			
Penicillin Allergy	91936005			
Urticaria	126485001			
Anemia	271737000	Anemia	271737000	
Anorexia	79890006	Anorexia	79890006	*Refers to low appetite of any cause*
Arthralgia	57676002	Arthralgia	57676002	
Back Pain	161891005			
Elbow Pain	74323005			
Finger Pain	18876004			
Hip Pain	49218002			
Knee Pain	30989003			
Leg Pain	156345110001191 08			
Limb Pain	90834002			
Noise from Sternum	250119000			
Toe Pain	285365001			
Asplenia	51242008	Asplenia	51242008	*Post-splenectomy patients automatically assumed to have this condition*
Asthma	195967001	Asthma	195967001	
Ataxia	20262006	Ataxia	20262006	
Motor Ataxia	59546009			
Autoantibody Titer Negative	165878000	Autoantibody Titer Negative	165878000	
Autoantibody Titer Positive	165879008			
Positive ANA Antibody Titer	444551008			
Positive ANA titer	165850001			
Positive LKM Antibody	720330001.POS			
Positive LKM titer	720330001.POS			
Positive PANCA Titer	401078001.POS			
Benign Tumor	3898006	Benign Tumor	3898006	
Breast Papilloma	99571000119102			
Bone Fracture	125605004	Bone Fracture	125605004	*Broken tooth is included because it*
Fracture of body of vertebra	445734009			

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Rib Fracture	45910007			*is a fracture of bone tissue*
Broken Tooth	36202009			
Bone Pain	12584003	Bone Pain	12584003	
Bruising	125667009	Bruising	125667009	
Bruising Diathesis	424131007			
Cancer	363346000	Cancer	363346000	
Cervical Cancer	363354003			
Cataracts	193570009	Cataracts	193570009	
Advanced Cirrhosis	123717006	Cirrhosis	19943007	
Cirrhosis	19943007			
Aphasia	87486003	Cognitive Impairment NOS	386806002	*Includes mild or perceived cognitive impairment*
Cognitive Impairment NOS	386806002			
Ankylosing Spondylitis	9631008	Comorbid Autoimmune Disease	85828009	*Excludes Diabetes Mellitus Type 1*
Autoimmune gastritis	84568007			
Autoimmune Urticaria	402397006			
Celiac	396331005			
Comorbid Autoimmune Disease	85828009			
Hashimoto Thyroiditis	21983002			
Lupus	200936003			
Psoriatic Arthritis	156370009			
Rheumatoid Arthritis	69896004			
Sjorgen Syndrome	83901003			
Idiopathic Thrombocytic Purpura	32273002			
Cough	49727002	Cough	49727002	
Dark Urine	720001001	Dark Urine	720001001	
Degenerative Disc Disease	77547008	Degenerative Disc Disease	77547008	
Herniated Disc	2304001			
Dental Caries	80967001	Dental Disorder	234947003	*Excludes broken tooth, which is considered a bone fracture*
Dental Disorder	234947003			
Diabetes Mellitus NOS	73211009	Diabetes Mellitus NOS	73211009	

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Diabetes Mellitus Type I	46635009			
Disorder of Pancreas	3855007	Disorder of Pancreas	3855007	
Pancreatic Cyst	31258000			
Pleural Effusion	60046008	Disorder of Pleura	3855007	
Droopy Eyelid	11934000	Droopy Eyelid	11934000	
Dysglycemia	166922008	Dysglycemia	166922008	*Excludes declared Diabetes Mellitus*
Hypoglycemia	302866003			
Edema	267038008	Edema	267038008	
Ehlers Danlos Syndrome	398114001	Ehlers Danlos Syndrome	398114001	*All subtypes*
Adrenal Insufficiency	386584007	Endocrinopathy	362969004	*Excludes Hashimoto (autoimmune) thyroiditis*
Endocrinopathy	362969004			
Eruption of Skin	271807003	Eruption of Skin	271807003	*Excludes acne*
Pityriasis Rosea	77252004			
Failure to Thrive	54840006	Failure to Thrive	54840006	
Fatigue	84229001	Fatigue	84229001	
Fever	386661006	Fever	386661006	
Fibromyalgia	203082005	Fibromyalgia	203082005	
Cholecystectomy	38102005	Gall Bladder Problem	28231008. PROB	*SNOMED has no code for generic gallbladder problem*
Cholestasis	33688009			
Gall Bladder Problem	28231008.PROB			
Gallstone	2359190008			
Gastric Ulcer	429040005	Gastric Ulcer	429040005	
Bloating Symptom	248490000	Gastrointestinal Upset	162059005	*Excludes abdominal pain with no assumed cause*
Gastritis	4556007			
Gastrointestinal Upset	162059005			
GERD	235595009			
Nausea	422587007			
Vomiting	422400008			
Gilbert Syndrome	27503000	Gilbert Syndrome	27503000	
Gingivitis	66383009	Gingivitis	66383009	
Gluten Allergy Non-Celiac	414285001	Gluten Sensitivity	441831003	*Excludes autoimmune celiac disease*
Gluten Sensitivity	441831003			
Gynecomastia	4754008	Gynecomastia	4754008	

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Cluster Headache	193031009	Headache	25064002	
Headache	25064002			
Headache Chronic	431237007			
Migraines	37796009			
Heart Disease	56265001	Heart Disease	56265001	
Myocarditis	50920009			
Hepatic Encephalopathy	13920009	Hepatic Encephalopathy	13920009	
Acute Hepatic Failure	19720009	Hepatic Failure	59927004	
Hepatic Failure	59927004			
Hyperammonemia	9360008	Hyperammonemia	9360008	
Hypercholesteremia	13644009	Hypercholesteremia	13644009	
Hyperhidrosis	312230002	Hyperhidrosis	312230002	
Hypoferritinemia	165623008	Hypoferritinemia	165623008	
Hypotension	45007003	Hypotension	45007003	
Hysterectomy	236886002	Hysterectomy	236886002	
Common Cold	82272006	Infectious Disease	40733004	
Dientamoeba fragilis Infection	240367005			
Fungal Skin Infection	14560005			
Helicobacter pylori Infection	721730009			
Infectious Disease	40733004			
Infectious Mononucleosis	271558008			
Influenza	6142004			
MRSA Infection	266096002			
Mycosis	3218000			
Otitis	43275000			
Pneumonia	233604007			
Sepsis	91302008			
Sinusitis	39671009			
Skin Infection	108365000			
Urinary Tract Infection	68566005			
Varicella Zoster Infection	38907003			
Viral Hepatitis	3738000			
Insomnia	193462001	Insomnia	193462001	

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Irritable Bowel Syndrome	10743008	Irritable Bowel Syndrome	10743008	
Jaundice	18165001	Jaundice	18165001	
Injury to Anterior Crucial Ligament	127292004	Joint Injury	125610000	*Includes all sprains and strains*
Joint Injury	125610000			
Kerato-conjunctivitis Sicca	302896008	Kerato-conjunctivitis Sicca	302896008	
Kidney Stones	95570007	Kidney Stones	95570007	
Knee injury	1256010008	Knee injury	1256010008	
Lactose Intolerance	2764325008	Lactose Intolerance	2764325008	
Leukopenia	419188005	Cytopenia	508200005	
Lymphopenia	48813009			
Thrombocytopenia	302215000			
Hepatic Fibrosis	62484002	Liver Damage NOS	243978007	*Used when these types of liver damage are reported but by unknown cause. Typically assumed AIH-induced*
Hepatomegaly	80515008			
Liver Cyst	85057007			
Liver Damage NOS	243978007			
Mole	400096001	Mole	400096001	
Anxiety	48694002	Mood Issue (Dysphoric Mood)	271596009	*Includes any hint of negatively altered mood; not just diagnosed psychiatric issues*
Depressed Mood	366979004			
Dysphoric Mood	30819006			
Feeling Stressed	224974006			
Labile Mood	18963009			
Mood Issue	271596009			
Panic Attack	225624000			
Reactive Depression	87414006			
Muscle Twitch	60238002	Muscle Twitch	60238002	
Muscle Spasm	221360009	Myalgia	68962001	
Myalgia	68962001			
NAFLD	197315008	NAFLD	197315008	
Bilirubin Normal	166611006	Normal LFT	250119000	
Stabilized AIH	408335007.STAB			
Stabilized AIH Serum	408335007.STAB.SER			
Osteoarthritis	396275006	Osteoarthritis	396275006	
Osteopenia	312894000	Osteopenia	312894000	
Ovarian Cyst	79883001	Ovarian Cyst	79883001	

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Ruptured Ovarian Cyst	95598005			
Numbness of Lower Limb	309537005	Paresthesia	91019004	
Paresthesia	91019004			
Primary Biliary Cirrhosis	31712002	Primary Biliary Cirrhosis	31712002	
Physical Pain	22253000	Physical Pain	22253000	*Coded only when no further nature of the pain is known.*
Lung Collapse	46621007	Pleural Disorder	77252004	
Polymyositis	31384009	Polymyositis	31384009	
Portal Hypertension	34742003	Portal Hypertension	34742003	
Positive Lyme Titer	310567000	Positive Lyme Titer	310567000	*Not included with other antibody titers due to lack of hepatic involvement*
Caesarean Section	11466000	Pregnancy	289908002	*A user with history of Caesarean section is assumed to have been pregnant.*
Pregnancy	289908002			
Primary Sclerosing Cholangitis	197441003	Primary Sclerosing Cholangitis	197441003	
Pruritus	418363000	Pruritus	418363000	
Raynaud Phenomenon	266261006	Raynaud Phenomenon	266261006	
Infertility	6738008	Reproductive Disorder	362968007	*Excludes menstrual issues*
Polyuria	28442001			
Reproductive Disorder	362968007			
Odynypnea	75483001	Respiratory Disorder	50043002	*Excludes respiratory infections, which are considered infectious diseases*
Respiratory Disorder	50043002			
Rhabdomyolysis	240131006	Rhabdomyolysis	240131006	
Splenomegaly	16294009	Splenomegaly	16294009	
Asymptomatic AIH	408335007. ASYMP	Stabilized AIH	408335007. STAB	*Excludes AIH known to be LFT-stable (see normal*
Compensated AIH	408335007.COMP			

Issue	SNOMED	Parent Issue	Parent SNOMED	*Parent Issue Comment*
Stabilized AIH Antibody	408335007.STAB. AB			*LFT) and symptomatically stable AIH*
Stabilized AIH Gross	408335007.STAB. BIOP			
Constipation	14760008	Stool Abnormal	201222006	
Stroke	230690007	Stroke	230690007	
Tachycardia	3424008	Tachycardia	3424008	
Hypothyroidism	40930008	Thyroiditis	82119001	*Excludes autoimmune thyroiditis*
Thyroid Disorder NOS	82119001			
Type 2 AIH	721712002	Type 2 AIH	721712002	
Esophageal Varices	28670008	Varices	128060009	
Gastric Varices	91109007			
Varices	128060009			
Vertigo	399153001	Vertigo	399153001	
Vision Disorder	95677002	Vision Disorder	95677002	*Only includes anomalies correctable with optometric means*
Vitiligo	56727007	Vitiligo	56727007	
Moon Facies	67967009	Weight Gain	8943002	*Moon facies are assumed to be a weight gain issue*
Weight Gain	8943002			
Weight Loss	89362005	Weight loss	89362005	
Xerostomia	87715008	Xerostomia	87715008	

APPENDIX V. PAIN & INJURY REGEX MEMBER DICTIONARY

Negations
"never", "do_not", "don_t", "you", "i_wonder", "heard", "wondered", "understand", "not_had", "nt_had", "n_t_had", "i_think", "was_thinking"

Pain/Injury Type: General Physical Pain
"pain","hurt", "ache", "aching", "troubl", "cramp", "sting", "tear", "issue","sore", "feeling_in", "felt_my"

Pain/Injury Type: Neuropathic Pain
"_burn", "tingl", "_sting", "paresth"

Pain/Injury Type: Fracture
"broke", "fractur", "break", "snap", "shatter"

Pain/Injury Type: Soft Tissue Injury
"sprain", "strain", "tear", "tore", "bruis", "torn", "dislocat"

Pain/Injury Type: Muscle Pain & Injury
"sore", "cramp", "spasm", "knot", "tight", "spastic"

Location: Joint (non-Hip)
"cartilage","joint","tendon","tendinitis","knee","meniscus","meniscal", "_acl_","knuckle","thumb","ankle","heel","shoulder","elbow","bursitis", "bursa"

Location: Hip
pelvis,_hip_","_hips_","pelvic"

Location: Lower Limb
"toe","leg","thigh","femur","_shin_","tibula","fibula","calf","calves", "foot","feet"

Location: Upper Limb
"arm","humerus","hand","finger","thumb","pinky","pinkie"

Location: Back & Spine
"_back_","lumbar","spinal","neck","thorac","spine","sacral","tailbone", "verteb","disc"

Location: Oral & Dental
"tooth","teeth","gum","mouth","oral","tongue"

Location: Abdomen

"tummy”,”stomach”,”gut_”,”llq”,”luq”,”ruq”,”rlq”,”gastro”,”abdom”,”quadrant”,”kidney”,”pancrea”,”bowel”,”colon”,”midsection”,”mid_section”,”belly”,”navel”,”spleen”,”pancrea”,”urq”,”ulq”,”lrq”,”flank”,”left_lower”,”left_upper”,”right_lower”,”right_upper”,”lower_right”,”lower_left”,”upper_right”,”upper_left"

Location: Thorax

"thora”,”chest”,”lung”,”_rib_”,”_ribs_”,”ribcage"

Location: Muscle (NOS)

"_myalg”,”muscl”,”muscul"

Location: Bone (NOS)

"bone”,”bony"

Location: Head (Cranium)

"head”,”skull”,”eye”,”face”,”facial”,”migrain”,”bell_"

APPENDIX VI. REGEX MEMBERS FOR DETECTION OF NON-PAIN RELATED SIGNS, SYMPTOMS, AND COMORBIDITIES

Note: This appendix only contains the distinguishing terms for experiencers of all noted signs, symptoms, and comorbidities.

Negations

```
"dont", "do_not", "don_t", "has_not", "hasnt", "hasn_t", "has_nt",
"never_had", "never_have", "negative", "normal", "doesnt", "does_not",
"doesn_t"
```

Experiencers

```
"has", "dx", "having", "feeling", "suffer", "diagnosed", "becoming",
"getting", "i_am", "i_m", "_im_", "any_one", "anyone", "any_body",
"anybody","also", "is_", "have", "my_", "his", "her", "sons", "son_s",
"hes", "he_s", "he_is", "for_", "going_through", "get_", "dx"
```

APPENDIX VII. LAB TEST REGEX MEMBER DICTIONARY

Test Type Indicators

Test Type: Liver Function Tests (LFTs)

```
"_ast_","_alt_","astalt","altast","spt","sgot","liver_function","liver_
test","liver_lab","liver_level","hepatic_level","enzyme","livers_","lft
","hft","ggt","alkaline","alp_","phosphatase","_asts","_alts","_bun_","
nitrogen","numbers","creatinin","altsasts","astsalts","alpast","alpalt"
,"astalp","altalp","ammonia","ammone","bili_","bilirub","billirub","cre
atinin","creatine","cpk"
```

Test Type: White Blood Cell Tests (WBC)

```
"bloods","wbc","white_","whites_","count_","blood_cell","cell_count"
```

Test Type: Red Blood Cell Tests (RBC)

```
"red_blood","rbc","eryrthrocyte","iron","_fe_"
```

Test Type: AIH Autoantibodies (AABs)

```
"lkm","anca","_ana_","_anti_","_ab_","antibod","antianc","antiana"
```

Test Value Indicators

Value Indicators: Low Prefixes/Suffixes

```
"low_", "depress", "lower", "only", "drop", "fell"
```

Value Indicators: High Prefixes/Suffixes

```
"raise", "raising", "rise", "rising", "elevate", "high", "nonresp",
"non_resp", "jack"
```

Value Indicators: Normal Prefixes/Suffixes

```
"_normal", "ok_", "okay", "good", "great", "perfect", "fine",
"no_problem", "negative", "_stable", "_stabili", "respond",
"wonderful","acceptab", "not_a_problem", "nt_a_problem", "not_problem",
"nt_problem", "to_normal", "back_normal", "improv", "better"
```

Value Indicators: Abnormal Prefixes/Suffixes

```
"bad", "abnormal", "not_good", "not_fine", "_off", "positive",
"unstable", "unstabil", "all_over", "horribl", "terribl", "awful",
"not_respond", "nt_respond", "problem", "issue", "wrong", "screw",
"not_normal", "worse"
```

APPENDIX VIII. CLASSIFIED SCHEDULE OF ALL DOMAINS & SUBDOMAINS OF USER COMMUNICATION

Domain	Name	Description & Scope Notes	N
DEMO	**Demographics**	**Commonly elicited social characteristics of an individual**	**(N/A)**
DEMO.AGE	Age	Chronological age-typically in years	14
DEMO.AGE.CGMIN	Caregiver For Minor	The user is posting on behalf of someone who is under the age of majority	1
DEMO.GEN.FEM	Female Gender	Affiliating with the female (woman or feminine) gender	1
DEMO.GEN.MALE	Male Gender	Appearing to affiliate as a male gender (man or masculine) individual	1
DEMO.MARR	Married	In an ideally permanent and monogamous relationship	1
ENV	**Physical Environment**	**The physical surroundings of a person, including geographical location and natural environment**	**(N/A)**
ENV.PET	Having pets	In possession of non-human animal companion	3
ENV.PROB	Problems in the physical environment	Problems with one's physical surroundings	1
ENV.RSLOC	Geographical residence location	Geographical residence location	1
ENV.RSLOC.AUS	Australian Resident	Lives in Australia	1
ENV.RSLOC.US	US Resident	Lives in the United States	11
ENV.WSMK	History of Wood Smoke Exposure	A history of wood smoke being in the person's surroundings	2
FAMHX	**Family History**	**Medically-recognized symptoms and diagnoses of blood relatives**	**(N/A)**
FAMHX.AI	Family History of Autoimmune Disease	Evidencing a blood relative with a disease where the immune system attacks the body	12
FAMHX.CA	Family History of Cancer	Evidencing a blood relative with a disease that involves an uncontrollable malignant growth of cells	1

Domain	Name	Description & Scope Notes	N
FAMHX.ID	Family History of Infectious Disease	Evidencing a blood relative that had a communicable disease	1
FAMHX.PSY	Family History of Psychological Issues	Evidencing a blood relative who had problems and-or diagnoses of mood issues	1
FIN	**Finances**	**Pertaining to monetary instruments, acquirement thereof (e.g., employment) and sureties (e.g., insurance)**	**(N/A)**
FIN.POS	Positive financial status	Favorable financial factors	(N/A)
FIN.POS.INS	Favorable Insurance Status	Favorable health insurance status	2
FIN.POS.INS.MEDS	Favorable Medication Insurance	In possession of what is believed to be a good medication coverage plan	1
FIN.PROB	Problems with finance	Unfavorable financial factors	(N/A)
FIN.PROB.INS	Insurance Problems	Having problems with insurance	3
FIN.PROB.INS.MEDS	Insurance Problems: Medications	Having problems getting medications due to insurance	13
FIN.PROB.MEDS	Problems Affording Medications	Having problems affording medications - Not attributable to insurance issues	1
GRSOC	**Support Group Socializing**	**Greeting others within context of the group discussion itself**	**7**
GRSOC.CONF	Support Group Annual Conference	Socialization relating to the annual group AIH conference	1
GRSOC.GRE	Giving greetings to group members	Simple greetings exchanged over the online support group	6
GRSOC.HUMOR	Group Humor	Humor shared over the online support group	1
GRSOC.INTRO	Introducing Oneself on the Group	A new user introducing themselves to the online support group	1
GRSOC.IRL	Group members in real life	Meeting or knowing group members in real life	1
GRSOC.PRIV	Privacy issues on group	Respecting a group member's privacy	1

Domain	Name	Description & Scope Notes	N
HCS	**Healthcare System**	**Matters pertaining to the user's interaction with healthcare, including clinic and clinicians-providers. Excludes personal medical story (MEDXS) and insurance (FIN.INS) discussions.**	**(N/A)**
HCS.POS	Healthcare System: Pros	Positive aspects of healthcare system	1
HCS.POS.CLIN	Positive Clinician Experiences	Having had positive experiences with clinicians-providers	5
HCS.POS.CLIN.COOP	Clinicians cooperating	The user's clinicians are in fact cooperating	2
HCS.POS.TECH	Benefits of Technology in Healthcare	Positive technological aspects of healthcare system	1
HCS.PROB	Healthcare System: Problems	Negative experiences with the healthcare system	1
HCS.PROB.CLIN	Problems with Clinicians	Problems with clinician	4
HCS.PROB.CLIN.COOP	Clinicians Not Cooperating	Problems with clinicians unable to cooperate	1
HCS.PROB.CLIN.DAGR	Disagreeing with Clinician	Disagreeing with a clinician	1
HCS.PROB.CLIN.IGN	Clinician Ignoring Patient	Problems with clinician ignoring patient	1
HCS.PROB.CLIN.INCMP	Clinician Being Incompetent	Problems with clinician being ignorant or incompetent	2
HCS.PROB.DXLAB	Problems getting a diagnosis	Problems and conflict arising from diagnosis and-or lab tests	4
HCS.PROB.GEO.MOVE	Problems with clinicians due to frequent moving	Problems with clinicians due to frequent moving	1
HCS.PROB.WAIT	Excessive Waiting for Healthcare	Problems having to wait for healthcare	8
HCS.PROB.WRHEP	Healthcare thinking wrong hepatitis	Problems with healthcare system believing that the user has viral or alcoholic hepatitis	9

Domain	Name	Description & Scope Notes	N
ISHR	**Sharing Information**	**Sharing information for academic purposes with no support intended. Indicates advanced knowledge of the subject matter, which is typically health.**	**1**
ISHR.CAM	Academic interest in CAM	Academic (not necessarily personal) interest in complementary-alternative medicine	1
ISHR.HCS	Academic discussion on healthcare system	Sharing factual information about the healthcare system-not as a form of support	2
ISHR.HLTH	Academic discussion on health and medical issues	Sharing factual information about health and medicine - not as a form of support	25
MEDXS	**Medical stories/histories**	**Medically-recognized elements of health, including diagnoses, medication intake, lab test results, signs, and symptoms.**	**19**
MEDXS.AI	Histories of autoimmune disease	Histories of autoimmune disease including AIH	27
MEDXS.DXLAB	Discussing lab results and diagnoses pertinent	Discussing lab results and diagnoses pertinent	61
MEDXS.GENE	Discussing genetic test results	Revealing one's genetic test results	1
MEDXS.HOSP	History of hospitalization	History of hospitalization	4
MEDXS.MEDS	Medication intake	Affirming medication intake	56
MEDXS.MEDS.COMP	Treatment compliance	Adhering to treatment and lifestyle recommendations given by the clinician or provider	2
MEDXS.MEDS.EFF	Medication effective after intake	Evidencing that a medication had a desired effect after intake	1
MEDXS.MEDS.FEAR	Fear of Medications	Expressing fear of medications	2
MEDXS.MEDS.INEFF	Medications Ineffective	Discussing medications that did not have their intended effect	3
MEDXS.MEDS.LOWRX	Lowering of drug dosages	Discussing having lowered medication dosage. Related to VPT.LOWRX.AGR	1

Domain	Name	Description & Scope Notes	N
MEDXS.MEDS.WOR	Worried about medication effects	Worried about medications - usually side effects	1
MEDXS.NDC	Not Diagnosed but Concerned	Expressing worry due to concern about having a disease and not being diagnosed yet	1
MEDXS.PREG	Discussing one's pregnancy	Discussing one's pregnancy	3
MEDXS.PROC	Stories about procedures	Discussing procedures and surgery - excludes diagnostic procedures e.g. biopsy	6
MEDXS.SYMP	Medical symptoms and comorbidities	Discussion of symptoms - ADEs - and comorbid diagnoses	59
MEDXS.TRNSP	Liver Transplant History	Stories and histories about liver transplant	1
NCOMP	**Noncompliance**	**The act of not adhering to provider-mandated treatments**	**(N/A)**
NCOMP.MEDS	Medication Noncompliance	The act of not adhering to one's prescription drug regimen	4
QOLS	**Quality of Life Factors**	**Psychosocial factors ranging from emotions to coping to abilities to perform the daily activities to one's personal satisfaction**	**(N/A)**
QOLS.DECL	Declining QOL	Evidencing a declining quality of life (QOL)	(N/A)
QOLS.DECL.EMP	Declining Employment	Declining ability to pursue employment	1
QOLS.DECL.EXER	Declining Exercise Capability	Exercise capabilities declined or declining	3
QOLS.POS	Favorable QOL Factors	Favorable quality of life (QOL) factors	(N/A)
QOLS.POS.ADL	Favorable ADL Abilities	Favorable capabilities in performing activities of daily living (ADLs)	1
QOLS.POS.EMP	Favorable Employment Status	Discussing the ability to be employed	3
QOLS.POS.NRG	Favorable Energy Levels	Discussing having more physical energy. Only use if there is no exact QOL measure (ADL, EXER, etc.) stated	2
QOLS.PROB	Poor QOL	Evidencing a poor quality of life	4

Domain	Name	Description & Scope Notes	N
QOLS.PROB.ADL	Poor ADL Capabilities	Poor quality of life due to inability to do activities of daily living (ADLs)	6
QOLS.PROB.DIET	Inability to eat the desired diet	Quality of life impaired due to inability to eat preferred foods	1
QOLS.PROB.EMO	Emotional Issues interfering with QOL	Poor quality of life due to emotional problems. Only use if no other QOL (e.g. ADL) is affected	2
QOLS.PROB.EMP	Poor Employment Status	Poor quality of life due to lack of employment	7
QOLS.PROB.EMP.WRHEP	Employment discrimination: Wrong Hepatitis	Employment discrimination: Wrong Hepatitis	1
QOLS.PROB.EXER	Poor Exercise Ability	Poor quality of life due to inability to exercise	1
QOLS.PROB.HHD	Hyperhidrosis	QOL interfered with by excessive sweating	1
QOLS.PROB.NRG	Energy/fatigue problems	Energy-fatigue problems	2
QOLS.PROB.SEX	Sexual problems	Sexual problems	1
QOLS.PSY	Psychological QOL	Mental feelings and attitudes that affect life - Can be positive or negative	1
QOLS.PSY.ANTIC	Feeling anticipation	Expressing a feeling of looking forward to something	1
QOLS.PSY.CNFU	Confused attitude	Confused attitude	2
QOLS.PSY.DOU.MEDS	Doubting medical treatment	Skepticism at traditional medical treatments	1
QOLS.PSY.FEAR.AIH	Fear of AIH	Expressing fear of autoimmune hepatitis	1
QOLS.PSY.FEAR.DEATH	Fear of death	Expressing fear of death	1
QOLS.PSY.FEAR.HLTH	Fear for health	Being afraid for one's health in general	1
QOLS.PSY.FEAR.MEDS	Fear of medication	Being afraid of medications (usually due to adverse effects)	3
QOLS.PSY.FRUST.AIH	Frustrated With AIH	Becoming particularly upset due to AIH and its sequelae	4
QOLS.PSY.FRUST.AIS	Frustrated with informational/advice support	Becoming upset due to poor quality of information or advice	2
QOLS.PSY.FRUST.MEDS	Frustrated with Medications	Being dissatisfied with medications	2

Domain	Name	Description & Scope Notes	N
QOLS.PSY.FRUST.PLYRX	Frustration due to polypharmacy	Expressing frustration due to taking too many medications (polypharmacy)	1
QOLS.PSY.FRUST.PPAIN	Frustrated With Physical Pain	Becoming particularly upset due to pain	1
QOLS.PSY.FRUST.SUP	Complaining About Support	Complaining about support quality	5
QOLS.PSY.FRUST.WRHEP	Frustration Due to Wrong Hepatitis	Frustration due to public and-or professionals not understanding that the user's hepatitis is autoimmune - nonviral and nonalcoholic	2
QOLS.PSY.HOP	Expressing Hope	Expressng a positive sense that things will become better	1
QOLS.PSY.GRA.ADVOC	Grateful for advocacy efforts	Grateful for advocacy efforts in AIH	1
QOLS.PSY.GRA.FAM	Grateful for family	Grateful for relatives and their actions	1
QOLS.PSY.GRA.HLTH	Grateful for Good Health	Expressing thankfulness due to good health	3
QOLS.PSY.GRA.MEDS	Grateful for Medications	Expressing thankfulness due to medications working well	5
QOLS.PSY.GRA.PLM	Grateful for similar patients	Grateful for similar patients	4
QOLS.PSY.JOY	Joyous	Expressing a sense of happiness	3
QOLS.PSY.RLG	Religious	In belief of a dogmatic higher power and spirituality	1
QOLS.PSY.RSG	Resignation	Expressing a sense of giving up	5
QOLS.PSY.SAD	Sadness	Expressing solemn negative emotion	1
QOLS.PSY.WISH.CONF	Wishing to attend conference	Wanting to attend a conference (typically AIH related)	1
QOLS.PSY.WISH.HLTH	Wishful about health	Expressing the desire for one's health to improve	1
QOLS.PSY.WOR.DXLAB	Worried about diagnosis or lab results	Worried about diagnosis or lab results	1
QOLS.PSY.WOR.HLTH	Worried about health	Worried about health	2
RSPRT	**Research Participation**	**Participates as a subject in AIH-related research**	**4**

Domain	Name	Description & Scope Notes	N
SOC	**Social factors**	**Factors pertaining to the individuals who physically surround the user. Excludes marital status (DEMO.MARR). Excludes the provision and receiving of online support from other group members (SUP)**	**(N/A)**
SOC.HGNDK	Has Grandchildren	Has offspring who in turn have offspring	3
SOC.HHA	House Helper available	Someone else is available to help do chores in and around the house	1
SOC.HKIDS	Has kids	Has offspring (regardless of age)	7
SOC.HX.ABUSE	History of abuse	History of being a victim of physical-sexual-or emotional abuse	1
SOC.HX.LSUP	Social history of a lack of support	Social history of a lack of support	1
SOC.JOB.CLIN	Works as a clinician	Having a job as a nurse, doctor, or other medical provider	8
SOC.MARR	Married	Being in an ideally permanent and monogamous relationship with a romantic partner	3
SOC.PROB.SEX	Sexual problems	Sexual problems	1
SOC.RLPLM	Other similar patients in real life	Knowing a similar patient in real life not from the group	2
SOC.TRAVL	Travels	Engages in visiting locations other than one's place of residence	1
STX	**Self-treatment stories**	**Sharing stories about personal and self-care treatments, including diet and exercise. These treatments might or might not be at behest of a provider.**	**3**
STX.CAFF	Caffeine usage	Caffeine usage as self therapy	1
STX.CAM	Having CAM Treatment	Discussion of engaging in complementary medicine treatment	5
STX.CAM.CEASE	Stopping CAM treatment	Cessation of complementary-alternative treatments	2
STX.CAM.CON	Considering CAM Treatment	Contemplating starting a complementary alternative treatment	3
STX.CAM.EFF	Effective CAM Treatment	CAM Treatment associated with improvement	1

Domain	Name	Description & Scope Notes	N
STX.CAM.INTOL	Intolerance of CAM	The inability to tolerate a complementary-alternative therapy	1
STX.CHEMX	Chemical exposure (avoiding)	Avoiding exposure to environmental or foodstuff toxins for health purposes	3
STX.DIET	Using Dietary Treatments	Discussion of diet as a self-care treatment	12
STX.DIET.CON	Contemplating Dietary Intervention	Contemplating using a dietary intervention	1
STX.DIET.GF	Using Gluten-Free Diet as Treatment	Discussion of gluten-free diet as self-care treatment	2
STX.DIET.LCARB	Using Low-Carbohydrate Diet as Treatment	Using Low-Carbohydrate Diet as Treatment	3
STX.EXER	Using Exercise as Treatment	Discussion of exercise as a self-care treatment	3
STX.NOALC	Avoiding Alcohol	The avoidance of consumption of beverage ethanol	1
STX.VMIN	Taking vitamins and minerals	The act of self-care therapy with USDA recognized vitamins and minerals	3
SUP	**Support**	**Communications made to request or provide assistance from or to another individual(s)**	**(N/A)**
SUP.AIAC	Requesting advice-informational support from AC	Requesting advice-informational support from AC - specifically mentioning him	18
SUP.GRACK	Gratefully acknowledging Support	Expressing thankfulness due to support	51
SUP.OFF	Offering Support	The act of giving or offering support	5
SUP.OFF.ADVOC	Advocating the disorder	Spreading the news about the disorder across the Internet and in real life. *Not searched for in support algorithms.*	5
SUP.OFF.AIS	Offering Advice-Information (Assumed Health-Related)	Giving informational support or advice. Assumed to be health-related	82
SUP.OFF.EMO	Offering Emotional-Social Support	Giving emotional (social) support	47

Domain	Name	Description & Scope Notes	N
SUP.PINQ	Personal inquiry (used in support provision)	Asking someone about themselves	19
SUP.REQ.AIS	Asking for Advice-Information (Assumed Health-Related)	Asking for advice or information as a form of support. Assumed to be health-related	51
TECH	**Technological factors**	**The user's relationship with electronic technology**	**(N/A)**
TECH.FRUST	Frustration with technology	Expressing negative emotions towards technology	3
TECH.SMUSE	Use of social media	Interest and competency in using online social media (not just using it per se)	1
TECH.UNFAM	Unfamiliar with technology	Not familiar or comfortable with using technology	2
TECH.WLRN	Willingness to learn how to use a new technology	Willingness to learn how to use a new technology	1
VPT	**Viewpoint (opinion)**	**A personal opinion not necessarily grounded in fact.**	**(N/A)**
VPT.ADVOC	Expressing an opinion about advocacy efforts	Demonstrating an opinion about efforts to shed light about AIH	2
VPT.AGR	Agreeing Viewpoint	Expressing agreement	7
VPT.CAM	Viewpoint on CAM	An opinion about complementary-alternative medicine therapies	1
VPT.CAM.AGR	Agreement with CAM therapies	Agreeing with or being biased towards complementary-alternative therapies	2
VPT.CAM.DOU	Doubting CAM therapies	Expressing skepticism that a complementary-alternative therapy may not work	2
VPT.HLTH	Health-related viewpoint	A viewpoint about health-related concerns, excluding complementary-alternative medical therapies.	2
VPT.LOWRX.AGR	Agreeing or for lowering prescription medication dosage	Agreeing or for lowering prescription medication dosage	1
VPT.POL	Political viewpoint	Expressing a viewpoint on politics	(N/A)

Domain	Name	Description & Scope Notes	N
VPT.POL.HLTH	Health Politics Viewpoint	Expressing health-related political viewpoints	2
VPT.RLG	Religious viewpoint	Religious viewpoint	2
VPT.SUGG	Making a suggestion	The act of recommending something - not intended as support	1

APPENDIX IX. LIST OF SUPPORT DETECTION DICTIONARY TERMS

Requesting Advice & Informational Support (REQ.AIS)

"please", "does any", "need advice", "want advice", "what does", "can any", "will any", "please give","does any", "can i ", "should i ", "would i ", "could i ", "wanna know", "wanna find", "give me advice", "can my", "can we","should my", "must i ", "one tell", "do you know", "why is", "why does", "what is", "what does", "where are", "one tell", "one know", "body know", "to know", "can any", "will any", "can some", "will some", "would some", "would any", "why do","would you", "what is", "let me know", "lmk", "where are", "why is", "where is", "where can", "would my ", "will my ", "how can", "how should", "m clueless", "cant figure", "not figure", "nt understand", "not understand", "am i suppose","what am ", "why am ", "how am ", "where am ", "where are ", "where is ", "how do", "will he ", "will she ", "will they ","is going on", " does", "dae ", "anyone else", "any one else", "anybody else", "any body else", "need an idea", "need ideas","could you", "would you", "wish i knew", "im asking", "am asking", "how would", "how should", "in advance", "wonder if", "want to learn","wanna learn", "want to know", "need to know", "any one can", "any body can", "anyone can", "anybody can", "one know", "body know","someone can", "some one can", "someone know", "some one know", "can someone", "can some one", "can somebody", "can some body","would someone", "would some one", "would somebody", "would some body", "would any body", "would anybody", "need advice", "have any of","body tell", "you tell", "just askin", "wanted to know", "want to know", "wanted to see", "want to see", "was askin", "be helpful", "be of help","need help", "want help", "help me", "can you give", "can you help", "anybody here", "any of you", "have a question", "question here","appreciated", "will this", "a question", "what have", "what has", "does somebody", "does someone", "can anyone", "can anybody","tell me why", "tell me how", "tell me what", "tell me which", "has anyone", "has any one", "has anybody", " is it ", "did they","any sugg", "any idea", "will this", "what has ", "what have", "was looking", "need info", "one have info", "body have info", "you have info","hellp", "helpp", "hellpp", "helppp", "helllp", "wtf", "what on ", "what the ", "why on ", "how on ", "in advance"

Offering Advice & Informational Support (OFF.AIS)

"you should", "my opinion", "read this", "check this", "should try", "http", "i read", "www", "pubmed","i think you", "you tried", "might want", "might need", "may want", "may need", "might wanna", "may wanna","try this", "you need", " imo ", "this link", "try this", "try to", "see and", "see if you", "if you can", "you really", "you

```
honestly","you do not", "dont you", "do you need", "just sayin", "you
tried", "what you", "you want", "you might", "you should", "dude",
"i think you", "can be", "could be", "should be", "can help", "could
help", "should help", "need to", "might help","should help", "could
work", "might work", "should work", "bad idea", "i recommend", "did it
work", "any better","you tried", "try doing", "should try", "could
try", "what you", "can you"
```

Gratefully Acknowledging Support (SUP.GRACK)

```
"thank", " ty ", "tysm", "thx", "thnk", "thanx", "tyvm", "tysm", "ty
vm", "ty sm"
```

Asking Administering colleague (AC) for Advice (SUP.AIAC)

These terms have been censored because they are variants on the administering colleague (AC)'s name.

Offering Emotional/Social Support (OFF.EMO)

```
"hug", "prayin", "prayer", "be ok", "be alright", "be all right",
"sorry", "hearts", "best wishes", "wish you luck",
"sending many", "sending much", "aww", "sweetie", "honey", "hunny",
"hon ", "congrat", "about you", "good thoughts", "sorry", "babe", "you
thoughts", "best of luck", "hope it", "feel for", "with you","you can
do", "good luck", "best wish", "hope things", "wishing", "in my
thought", "in our thought", "i feel for", "i know how", "how it feels"}
```

Personal Inquiry in Provision of Support (PINQ)

```
"is your", "is his", "is her", "is he", "do you", "does he", "does
she", "did you", "do you","have you", "can you", "you on", "you
taking", "you take", "does your", "what is you", "what is his", "what
is her", "whats you","whats his", "whats her", "whats their", "do
your", "are your", "is your";
```

REFERENCES

1. S Maiella, A Rath, C Angin, F Mousson, and O Kremp, *Orphanet and its consortium: Where to find expert-validated information on rare diseases* Rev Neurol (Paris), 2013. **169**: p. S3-S8.
2. National Organization for Rare Disorders. *NORD Physician Guides*. n.d. - 2016 30 September 2016]; Available from: http://nordphysicianguides.org/.
3. BP Smith, *Challenges and opportunities in rare disease drug development.* Clin Pharmacol Ther, 2016. **100**(4): p. 312-314.
4. J Forman, D Taruscio, VA Liera, LA Barrera, TR Cote, C Edfjall, D Gavhed, ME Haffner, Y Nishimura, M Posada, E Tambuyzer, SC Groft, and JI Henter, *The need for worldwide policy and action plans for rare diseases.* Acta Paediatr, 2012. **101**(8): p. 805-807.
5. LA Barrera and GC Galindo, *Ethical aspects on rare diseases.* Adv Exp Med Biol, 2010(686): p. 493-511.
6. NK Gatselis, K Zachou, GK Koukoulis, and GN Dalekos, *Autoimmune hepatitis, one disease with many faces: Etiopathogenetic, clinico-laboratory and histology characteristics.* World J Gastroenterology, 2015. **21**(1): p. 60-83.
7. M Aramayones, S Requena, B Gomez-Zuniga, M Pousada, and AM Banon, *The use of Facebook in Spanish associations of rare diseases: How and what is it used for?* Gaceta Sanitaria, 2015. **29**(5): p. 335-340.
8. YA Puius, LM Dove, DG Brust, DP Shah, and JH Lefkowitch, *Three cases of autoimmune hepatitis in HIV-infected patients.* J Clin Gastroenterology, 2008. **42**(4): p. 425-429.
9. GW Wong and MA Heneghan, *Association of extrahepatic manifestations with autoimmune hepatitis.* Digestive Dis, 2015. **2015**(33): p. Suppl 2:25-35.
10. U Iqbal, A Chaudhary, MA Karim, MA Siddiqui, H ANwar, and N Merrell, *Association of autoimmune hepatitis and celiac disease: Role of gluten-free diet in reversing liver dysfunction.* J Investig Med High Impact Case Rep, 2017. **5**(12): p. 2.
11. W Jo, YS Suh, SI Lee, YH Cheon, J Hong, SS Lee, JE Kim, GH Ko, and HO Kim, *Development of autoimmune hepatitis in a psoriasis patient without immunnosuppressive therapy.* Clin MOle Hepatology, 2017. **2017**(May 8).
12. A Czaja, *Factoring the intestinal microbiome into the pathogenesis of autoimmune hepatitis.* World J Gastroenterology, 2016. **22**(42): p. 9257-9278.
13. J Qian, Z Meng, J Guan, Z Zhang, and Y Wang, *Expression and roles of TIPE2 in autoimmune hepatitis.* Exp Ther MEd, 2017. **13**(3): p. 942-948.
14. YS De Boer, NM van Gerven, A Zwiers, BJ verwer, B van hoek, KJ van Erpecum, U Beurs, HR van Buuren, JP Drenth, JW den Ouden, RC Verdonk, GH Koek, JT Brouwer, MM Guichelaar, JM Vrolijk, G Kraal, CJ Mulder, CM Van Nieuwkerk, J Fischer, T Berg, F Stickel, C Sarrazin, C Schramm, AW Lohse, C Weiler-Norman, MM Lerch, M Nauck, H Volzke, G Hormuth, E Bloemena, HW Verspaget, V Kumar, A Zhernakova, C Wijmenga, L Franke, and G Bouma, *Genome-wide association study identifies variants associated with autoimmune hepatitis type I.* Gastroenterology, 2014. **2014**(147): p. 443-452.
15. EA Smuckler, *Iatrogenic disease, drug metabolism, and cell injury: Lethal synthesis in man.* Fed Proc, 1977. **36**(5): p. 1708-1714.

16. JH Ngu, RB Gearry, CM Frampton, and CA Stedman, *Autoimmune hepatitis: The role of environmental risk factors: A population-based study.* Hepatology Int, 2013. **7**(3): p. 869-875.
17. L Grasset, C Guy, and M Ollagnier, *Cyclines and acne: Pay attention to adverse drug reactions! A recent literature review.* Rev Med Interne, 2003. **2003**(24): p. 305-316.
18. W Davies, *Insights into rare diseases from social media surveys.* Orphanet J Rare Dis, 2016(2016;11): p. 151.
19. SG John, J Thorn, and R Sobonya, *Statins as a potential risk factor for autoimmune diseases: A case report and review.* Am J Ther, 2014. **21**(4): p. e94-e96.
20. V Alla, J Abraham, J Siddiqui, D Raina, GY Wu, NP Chalasani, and HL Bonkovsky, *Autoimmune hepatitis triggered by statins.* J Clin Gastroenterology, 2006. **40**(8): p. 757-761.
21. E Bjornsson, J Talwakar, S Treeprasertsuk, PS Kamath, N Takahashi, S Sanderson, M Neuhauser, and K Lindor, *Drug-induced autoimmune hepatitis: Clinical characteristics and prognosis.* Hepatology, 2010. **51**(6): p. 2040-2048.
22. ES Bjornsson, *Hepatotoxicity by drugs: The most common implicated agents.* Int J Mol Sci, 2016. **17**(2): p. 224.
23. SM Cohen, AM O'Connor, J Hart, NH Merel, and HS Te, *Autoimmune hepatitis associated with the use of black cohosh: A case study.* Menopause, 2004(2004;11): p. 575-577.
24. KM Gilbert, B Przybyla, NR Pumford, T Han, J Fuscoe, LK Schnackenberg, RD Holland, JC Doss, LA Macmillan-Crow, and SJ Blossom, *Delineating liver events in trichloroethylene-induced autoimmune hepatitis.* Chem Res Toxicol, 2009. **22**(4): p. 626-632.
25. SM Zhu, XF Ren, JX Wan, and ZL Xia, *Evaluation in vinyl chloride monomer-exposed workers and the relationship between liver lesions and gene polymorphisms of metabolic enzymes.* World J Gastroenterology, 2005. **11**(37): p. 5821-5827.
26. M Sebode, L Schulz, and AW Lohse, *Autoimmune(-like) drug and herb induced liver injury: New insights into molecular pathogenesis.* Int J Mol Sci, 2017. **18**(9): p. 1954.
27. F Alvarez, PA Berg, FB Bianchi, L Bianchi, AK Burroughs, EL Cancado, RW Chapman, WG Cooksley, AJ Czaja, and VJ Desmet, *International Autoimmune Hepatitis Group Report: Review of criteria for diagnosis of autoimmune hepatitis.* J Hepatol, 1999. **1999**(31): p. 929-938.
28. A Castiella, E Zapata, MI Lucena, and RJ Andrade, *Drug-induced autoimmune liver disease: A diagnostic dilemma of an increasingly reported disease.* Word J Hepatol, 2014. **6**(4): p. 160-168.
29. JPV Griend, *Common Polypharmacy Pitfalls*, in *Pharmacy Times*. 2009.
30. D Gleeson and MA Heneghan, *British Society of Gastroenterology (BSG) guidelines for management of autoimmune hepatitis.* Gut, 2011(2011;60): p. 1611-1612.
31. MR Lucey, *Clinical presentation and natural history of autoimmune hepatitis.* Clinical Liver Disease, 2014. **10 February 2014**.
32. P Bager, *The assessment and care of patients with hepatic encephalopathy.* British Journal of Nursing, 2017. **26**(13): p. 724-729.
33. X Fan, Y Zhu, R Men, M Wen, Y Shen, C Lu, and L Yang, *Efficacy and safety of immunosuppressive therapy for PBC-AIC overlap syndrome accompanied by decompensated cirrhosis: A real-world study.* Canadian Journal of Gastroenterology and Hepatology, 2018. **2018**(Aug 2): p. 1965492.
34. A Czaja, *Diagnosis and management of the overlap syndromes of autoimmune hepatitis.* Canadian Journal of Gastroenterology, 2013. **27**(7): p. 417-423.

35. MA Baven-Pronk, M Biewenga, JJ Van Silfhout, AP Van Denberg, HR Van Buuren, BJ Verwer, CM Van Niuewkerk, G Bouma, and B Van Hoek, *Role of age in presentation, response to therapy and outcome of autoimmune hepatitis.* Clinical Translational Gastroenterology, 2018. **9**(6): p. 165.
36. JK Karp, EK Akpek, and RA Anders, *Autoimmune hepatitis in patients with primary Sjorgen's syndrome: A series of two-hundred and two patients.* International Journal of Clinical Experimental Pathology, 2010. **3**(6): p. 582-586.
37. A Mansour, MR Mohajeri-Tehrani, M Samadi, H Gerami, M Qorbani, N Bellissimo, H Poutschi, and A Hekmatdoost, *Risk factors for non-alcoholic fatty liver disease-associated hepatic fibrosis in type 2 diabetes patients.* Acta Diabetology, 2019. **2019**(Jun).
38. M Hu, F Phan, O Bourron, P Ferre, and F Foufelle, *Steatosis and NASH in type 2 diabetes.* Biochimie, 2017. **2017 Dec**(143): p. 37-41.
39. VK Pandey, A Mathur, MF Khan, and P Kakkar, *Activation of PERK-elF2a-ATF4 pathway contributes to diabetic hepatotoxicity: Attenuation of ER stress by Morin.* Cell Signalling, 2019. **2019 Jul**(59): p. 41-52.
40. HC Jeffrey, MK Braitch, and C Bagnall, *Changes in natural killer cells and exhausted memory regulatory T cells with corticosteroid therapy in autoimmune hepatitis.* Hepatology Communications, 2018. **2**(4): p. 421-436.
41. LL Wong, HF Fisher, DD Stocken, S Rice, A Khanna, MA Heneghan, YH Oo, G Mells, S Kendrick, JK Dyson, and DE Jones, *The impact of autoimmune hepatitis and its treatment on health utility.* Hepatology, 2018. **2018**(4): p. e0.
42. Mayo Clinic. *Prednisone and other corticosteroids: Weigh the benefits and risks*. 2016 [cited 2018 August 14]; Available from: https://www.mayoclinic.org/steroids/art-20045692?pg=2.
43. A Kulanthaivel, JS Patel, CS Lammert, DJ Wild, S Milojevic, and JF Jones, *Social Media Content for Estimating Disease Factors in Otherwise Unreachable Patients: An Autoimmune Hepatitis (AIH) Case Study.* Submitting To: J Med Internet Res, 2020.
44. E Mahe, O Meyer, V Descamps, C Picard-Dahan, and B Crickx, *Early, severe and transient arthralgia induced by mycophenolate mofetil in a patient with erythrodermal psoriasis.* Annals of Dermatology and Venerology, 2002. **129**(8-9): p. 1054-1055.
45. AV Hassan, MD SInha, and S Waller, *A single-centre retrospective study of the safety and efficacy of mycophenolate mofetil in children and adolescents with nephrotic syndrome.* Clinical Kidney Journal, 2013. **6**(4): p. 384-389.
46. W Fukushima, M Fujioka, T Kubo, A Tamakoshi, M Nagai, and Y Hirota, *Nationwide epidemiologic survey of idiopathic osteonecrosis of the femoral head.* Clinical Orthopedic Related Research, 2010. **468**(10): p. 2715-2724.
47. D Nango, H Nakashima, Y Hirose, M Shiina, and H Ezichen, *Causal relationship between acute pancreatitis and methylprednisolone pulse therapy for fulminant autoimmune hepatitis: A case report and review of literature.* J Pharm Health Care Sci, 2018. **2018 May 31**(4): p. 14.
48. M Dirks, K Haag, H Pflugrad, AB Tryc, R Schuppner, H Wedemeyer, A Potthoff, HL Tillmann, K Sandorski, H Worthman, X Ding, and K Weissenborn, *Neuropsychiatric symptoms in hepatitis C patients resemble those of patients with autoimmune liver disease but are different from those in hepatitis B patients.* Journal of Viral Hepatology, 2018. **2018**(Aug 18).
49. A Aregay, M Dirks, V Schlaphoff, S Owusu Sekyere, K Haag, CS Falk, J Hengst, B Bremer, R Schuppner, MP Manns, H Pflugrad, M Cornberg, H Wedemeyer, and K Weissenborn, *Systemic inflammation and immune cell phenotypes are associated with*

neuro-psychiatric symptoms in patients with chronic inflammatory liver disease. Liver International, 2018. **38**(12): p. 2317-2328.
50. MK Janik, E Wunsch, J Raszeja-Wyszomirska, M Krawczyk, and P Milkiewicz, *Depression: An overlooked villain in autoimmune hepatitis?* Hepatology, 2019. **2019**(Feb).
51. LE Barry, J Sweeney, C ONeill, D Price, and LG Heaney, *The cost of systemic corticosteroid-indusced morbidity in severe asthma: A health economic analysis.* Respiratory Research, 2017. **18**(1): p. 129.
52. MA Sheiko, SS Sundaram, KE Capocelli, Z Pan, AM McCoy, and CL Mack, *Outcomes in pediatric autoimmune hepatitis and significance of azathioprine metabolites.* Journal of Pediatric Gastroenterology and Nutrition, 2017. **65**(1): p. 80-85.
53. AA Aljumah, H Al-Ashgar, H Fallatah, and A Albenmousa, *Acute onset autoimmune hepatitis: Clinical presentation and treatment outcomes.* Annals of Hepatology, 2019. **18**(3): p. 439-444.
54. G Perini, M Cotta Ramusino, E Sinforiani, S Bernini, R Petrachi, and A Costa, *Cognitive impairment in depression: Recent advances and novel treatments.* Neuropsychiatric Disease and Treatment, 2019. **2019 May 10**(15): p. 1249-1258.
55. Y Aizawa and A Hokari, *Autoimmune hepatitis: Current challenges and future prospects.* Clinical Experimental Gastroenterology, 2017. **2017**(10): p. 9-18.
56. M Sebode, C Weiler-Normann, T Liwinski, and C Schramm, *Autoantibodies in autoimmune liver disease - Clinical and diagnostic relevance.* Frontiers Immunology, 2018. **2018**(9): p. 609.
57. P Mueller, M Messmer, M Bayer, JM Pfeilschifter, E Hintermann, and U Christen, *Non-alcoholic fatty liver disease (NAFLD) potentiates autoimmune hepatitis in the CYP2D6 mouse model.* Journal of Autoimmunity, 2016. **2016 May**(69): p. 51-58.
58. T Himoto, K Fujita, T Nomura, J Tani, A Morishita, H Yoneyama, R Haba, and T Masaki, *Verification of B-lymphocyte activating factor's involvement in the exacerbation of insulin resistance as well as an autoimmune response in patients with nonalcoholic steatohepatitis and patients with HCV-related chronic liver disease.* Diabetology and Metabolic Syndromes, 2017. **2017 Jun 13**: p. 45.
59. AJ Czaja, *Epigenetic changes and their implications in autoimmune hepatitis.* European Journal of Clinical Investigation, 2018. **48**(4).
60. AJ Czaja, *Review article: Next-generation transformative advances in the pathogensis and management of autoimmune hepatitis.* Alimentary Pharmacological Therapy, 2017. **46**(10): p. 920-937.
61. TE Silva, G Colombo, and LL Schiavon, *Adiponectin: A multitasking player in the field of liver diseases.* Diabetes & Metabolism, 2014. **40**(2): p. 95-107.
62. KV Luong and LT Nguyen, *The role of vitamin D in autoimmune hepatitis.* Journal of Clinical Medical Research, 2013. **5**(6): p. 407-415.
63. NM Van Gerven, YS de Boer, CJ Mulder, CM van Nieuwkerk, and G Bouma, *Auto immune hepatitis.* Journal of Autoimmunity, 2016. **22**(19): p. 4651-4661.
64. NM Van Gerven, BJ Verwer, BI Witte, B Van Hoek, MJ Coenraad, and KJ Van Erpecum, *Dutch Autoimmune Hepatitis Working Group: Relapse is almost universal after withdrawal of immunosuppressive medication in patients with autoimmune hepatitis in remission.* Jourhal of Hepatology, 2013. **58**(1): p. 141-147.
65. B Hoeroldt, E McFarlane, A Dube, P Basumani, M Karajeh, MJ Campbell, and D Gleeson, *Long-term outcomes of patients with autoimmune hepatitis managed at a nontransplant center.* Gastroenterology, 2011. **140**(7): p. 1980-1989.

66. J De Luca-Johnson, KJ Wangensteen, J Hanson, E Krawitt, and R Wilcox, *Natural history of patients presenting with autoimmune hepatitis and coincident nonalcoholic fatty liver disease.* DIgestive Disease Science, 2016. **61**(9): p. 2710-2720.
67. RJ Holden, A Kulanthaivel, S Purkayastha, KM Goggins, and S Kripalani, *Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure.* Int J Med Inform, 2017. **2017 Dec**(108): p. 158-167.
68. JW Wen, MA Kohn, R Wong, M Somsouk, M Khalili, J Maher, and MM Tana, *Hospitalizations for autoimmune hepatitis disproportionately affect Black and Latino Americans.* American Journal of Gastroenterology, 2018. **113**: p. 243-253.
69. EV Golovanova, *Treatment of patients with overlap of primary biliary cirrhosis and autoimmune hepatitis.* Experimental Clinical Gastroenterology, 2010. **2011**(9): p. 140-148.
70. S Srivastava and JL Boyer, *Psychological stress is associated with relapse in type 1 autoimmune hepatitis.* Liver International, 2010. **30**(10): p. 1439-1447.
71. DS Smyk, EI Rigopoulou, L Muratori, AK Burroughs, and DP Bogdanos, *Smoking as a risk factor for autoimmune liver disease: What we can learn from primary biliary cirrhosis.* Annals of Hepatology, 2012. **11**(1): p. 7-14.
72. T Visseren and S Darwish Murad, *Recurrence of primary sclerosing cholangitis, primary biliary cholangitis and auto-immune hepatitis after liver transplantation.* Best Practices of Research in Clinical Gastroenterology, 2017. **31**(2): p. 187-198.
73. A Sonnenberg and WE Naugler, *Models of influence in chronic liver disease.* Liver International, 2010. **30**(5): p. 718-724.
74. GF Mells, A Kaser, and TH Karlsen, *Novel insights into autoimmune liver diseases provided by genome-wide association studies.* Journal of Autoimmunity, 2013. **2013 Oct**(46): p. 41-54.
75. MK Janik, E Wunsch, J Raszeja-Wyszomirska, M Moskwa, B Kruk, M Krawczyk, and P Milkiewicz, *Autoimmune hepatitis exerts a profound, negative effect on health-related quality of life.* Liver International, 2018. **11 September 2018**: p. 1-7.
76. Disease Natl Inst Diabetes Dig Kidney. *Drug Record: Herbal and Dietary Supplements*. LiverTox 2-13; Available from: https://livertox.nih.gov/Herbals_and_Dietary_Supplements.htm.
77. R Collier, *Patient engagement or social media marketing?* Canadian Medical Association Journal, 2014. **186**(8): p. E237-E238.
78. C Liddy, *Use of Facebook as part of a social media strategy for patient engagement.* Canadian Family Physician, 2017. **63**: p. 251-252.
79. R Rozenblum, F Greaves, and DW Bates, *The role of social media around patient experience and engagement.* BMJ Quality and Safety, 2017. **2017**(26): p. 845-848.
80. VK Dhar, Y Kim, JT Graff, AD Jung, J Garrett, LE Dick, J Harris, and SA Shah, *Benefit of social media on patient engagement and satisfaction: Results of a 9-month, qualitative pilot study using Facebook.* Surgery, 2018. **2018**(163): p. 565-570.
81. J Sohi, M Champagne, and S Shidler, *Improving health care professionals' collaboration to facilitate patient praticipation in decisions regarding life-prolonging care: An action research project.* Journal of Interprofessional Care, 2015. **29**(5): p. 409-414.
82. Inc. Facebook. *Facebook*. 2004-2017; Available from: http://www.facebook.com/.
83. M Duggan, NB Ellison, C Lampe, A Lenhart, and M Madden. *Demographics of key social networking platforms.* Pew Res Cent Internet Sci Tech 2015 2015; Available from: http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/
84. Twitter Inc. *Twitter Privacy Policy*. 2017; Available from: http://twitter.com/privacy?lang=en.

85. A Kulanthaivel, R Fogel, CS Lammert, and JF Jones, *Digital cohorts within the social mediome: An approach to circumventing common research challenges?* Clin Gastroenterology Hepatology, 2017. **15**(5): p. 614-618.
86. KR Schumacher, KA Stringer, JE Donohue, S Yu, A Shaver, RL Caruthers, BJ Zikmund-Fisher, C Fifer, C Goldberg, and MW Russell, *Social Media Methods for Studying Rare Diseases.* Pediatrics, 2014. **133**(5): p. e1345-e1353.
87. Statista: The Statistics Portal. *Statista: Number of monthly active facebook users worldwide*. n.d. - 2016 [cited 2016 24 November]; Available from: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.
88. SA Greeley, RN Naylor, LS Cook, SE Tucker, Lipton, and LH Philipson, *Creation of the web-based University of Chicago Monogenic Diabetes Registry: Using technology to facilitate longitudinal study of rare subtypes of diabetes.* J Diabetes Sci Technol, 2011. **5**(4): p. 879-886.
89. C Hawn, *Take two aspirin and tweet me in the morning: How Twitter, Facebook, and other social media are reshaping health care.* Health Affairs, 2009. **28**(2): p. 361-368.
90. K Wittmeier, C Holland, K Hobbs-Murrison, E Crawford, C Beauchamp, B Milne, M Morris, and R Keijzer, *Analysis of a parent-initiated social media campaign for Hirschsprung's disease.* J Med Internet Res, 2014. **16**(12): p. e288.
91. R Yu, *The Dr Pheo Blog and virtual counseling for rare diseases.* J Telemed Telecare, 2015. **21**(1): p. 54-57.
92. K Albright, T Walker, S Baird, L Eres, T Farnsworth, K Fier, D Kervitsky, M Korn, DJ Lederer, M McCormick, JF Steiner, T Vierzba, FS Wamboldt, and JJ Swigris, *Seeking and sharing: Why the pulmonary fibrosis community engages the Web 2.0 environment.* BMC Pulm Med, 2016. **16**: p. 4.
93. N Pemmaraju, Gupta, MA Thompson, and AA Lane, *Social media and Internet resources for patients with blastic plasmacytoid dendritic cell neoplasm (BPDCN).* Curr Hematol Malig Rep, 2016.
94. DE Winchester, D Baxter, MJ Markham, and RJ Beyth, *Quality of social media and web-based information regarding inappropriate nuclear cardiac stress testing and the Choosing Wisely campaign: A cross-sectional study.* Interact J Med Res, 2017. **2017 May 4**(1): p. e6.
95. K Tates, M Zwaanswijk, R Otten, S van Dulmen, PM Hoogerbrugge, WA Kamps, and JM Bensing, *Online focus groups as a tool to collect data in hard-to-include populations: Examples from paediatric oncology.* BMC Med Res Methodology, 2009(2009 Marc 3;9): p. 15.
96. S Kaufman and KA Whitehead, *Producing, ratifying, and resisting support in an online support forum.* Health (London), 2018. **22**(3): p. 223-239.
97. X Wang, K Zhao, and N Street, *Analyzing and predicting user participations in online health communities: A social support perspective.* J Med Internet Res, 2017. **19**(4): p. e130.
98. P Wicks, M Massagil, J Frost, C Brownstein, S Okun, T Vaughan, R Bradley, and J Heywood, *Sharing health data for better outcomes on PatientsLikeMe.* J Med Internet Res, 2010. **12**(2): p. e19.
99. H Alinia, S Moradi-Tuchayi, ME Farhangian, KE Huang, SL Taylor, S Kuo, I Richardson, and SR Feldman, *Rosacea patients seeking advice: Qualitative analysis of patients' posts on a rosacea support forum.* Journal of Dermatology Treatment, 2016. **27**(2): p. 99-102.
100. KK Walker, *A content analysis of cognitive and affective uses of patient support groups for rare and uncommon vascular diseases: Comparisons of May-Thurner, thoracic outlet,*

and superior mesenteric artery syndrome. Health Communications, 2015. **30**(9): p. 859-871.
101. MA Varga and TM Paulus, *Grieving online: Newcomers' constructions of grief in an online support group.* Death Studies, 2014. **38**(6-10): p. 443-449.
102. M Evans, L Donelle, and L Hume-Loveland, *Social support and online postpartum depression discussion groups: A content analysis.* Patient Education and Counseling, 2012. **87**(3): p. 405-410.
103. A Batenburg and E Das, *Emotional approach coping and the effects of online peer-led support group participation among patients with breast cancer: A longitudinal study.* Journal of Medical Internet Research, 2014. **16**(11): p. e256.
104. SJ Lepore, JS Buzaglo, MA Lieberman, M Golant, JR Greener, and A Davey, *Comparing standard versus prosocial internet support groups for patients with breast cancer: A randomized controlled trial of the helper therapy principle.* Journal of Clinical Oncology, 2014. **32**(36): p. 4081-4086.
105. L Chen and J Shi, *Social support exchanges in a social media community for people living with HIV/AIDS in China.* AIDS Care, 2015. **27**(6): p. 693-696.
106. SE Meredith, MJ Grabinski, and J Dallery, *Internet-based group contingency management to promote abstinence from cigarette smoking: A feasibility study.* Drug and Alcohol Dependency, 2011. **118**(1): p. 23-30.
107. SM Vambheim, SC Wangberg, JA Johnsen, and R Wynn, *Language use in an internet support group for smoking cessation: Development of sense of community.* Inform Health Soc Care, 2013. **38**(1): p. 67-78.
108. Facebook Inc. *Automated Data Collection Terms*. 2010-2018; Available from: https://www.facebook.com/apps/site_scraping_tos_terms.php.
109. Facebook Inc. *Graph API*. 2018-n.d.; Available from: https://developers.facebook.com/docs/graph-api/.
110. University Trustees of Indiana. *Research Using Online Tools and Mobile Devices*. n.d.-2018 [cited 2018 March 17]; Available from: http://researchcompliance.iu.edu/hso/hs_online_mobile.html.
111. S Golder, A Shahd, G Norman, and A Booth, *Attitudes towards the ethics of research using social media: A systematic review.* J Med Internet Res, 2017. **19**(6): p. e195.
112. S Alim, *An initial exploration of ethical research practices regardding automated data extraction from online social media user profiles.* First Monday, 2014. **19**(7): p. 105-127.
113. J Hudson and A Bruckman, *'Go Away': Participant objections to being studied and the ethics of chatroom research.* Inform Soc, 2004. **2004**(20): p. 127-139.
114. A Kulanthaivel, CS Lammert, and JF Jones, *A novel approach using social media to investigate patient-centric data in autoimmune hepatitis.*, in *Digestive Diseases Week*. 2018: Washington, DC.
115. Syncro Soft SRL, *Oxygen XML Editor*. n.d.-2016, Synchro Soft SRL: Craiova, Romania.
116. University of Waikato (NZ), *Weka: Waikato Environment for Knowledge Analysis*. 2018, University of Waikato (NZ): Hamilton, New Zealand.
117. A Kulanthaivel, CS Lammert, S Milojevic, DJ Wild, and JF Jones, *Facebook for population health research: A rare disease proof-of-concept.* Proc Am Med Inform Assoc Annu Symp, 2018.
118. Cytoscape.org, *Cytoscape.JS*. 1991-2017, Cytoscape.org: San Diego, CA.
119. AK McCallum, *MALLET: Machine Learning for Language Toolkit*. 2002-2016, University of Massachusetts: Amherst, MA.
120. JF Jones, M Pradhan, M Hosseini, A Kulanthaivel, and M Hosseini, *Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum.* JMIR Medical Informatics, 2018. **6**(4): p. e45.

121. MP Manns, AJ Czaja, JD Gorham, EL Krawitt, G Mieli-Vergani, D Vergani, and JM Vierling, *Diagnosis and management of autoimmune hepatitis.* Hepatology, 2010. **51**(6): p. 2193-2213.
122. BH Kim, HY Choi, M Ki, KA Kim, ES Jang, and SH Jeong, *Population-based prevalence, incidence, and disease burden of autoimmune hepatitis in South Korea.* PLoS One, 2017. **12**(8): p. e0182391.
123. WS Chou, YM Hunt, EB Beckjord, RP Moser, and BW Hesse, *Social media use in the United States: Implications for health communication.* J Med Internet Res, 2009. **11**(4): p. e48.
124. T Baldwin, P Cook, M Lui, A MacKinlay, and L Wang, *How noisy social media text, how diffrnt social media sources?* Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013: p. 356-364.
125. National Library of Medicine (United States). *RxNorm Overview*. 2005-2018; Available from: https://www.nlm.nih.gov/research/umls/rxnorm/overview.html.
126. M Marcus, *New trends in natural language processing: Statistical natural language processing.* Proc National Academy of Sciences, 1995. **92**(22): p. 10052-10059.
127. N Alvaro, Y Miyao, and N Collier, *TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations.* JMIR Public Health Surveillance, 2017. **3**(2): p. e24.
128. A Jimeno-Yepes, A MacKinlay, B Han, and Q Chen, *Identifying diseases, drugs, and symptoms in Twitter*, in *MEDINFO 2015: eHealth-enabled Health*. 2015.
129. F Altschul, W Gish, W Miller, EW Meyers, and DJ Lipman, *Basic Local Alignment Search Tool.* J Mol Biol, 1990. **1990**(215): p. 403-410.
130. AZ Klein, A Sarker, M Rouhizadeh, K O'Connor, and G Gonzalez, *Detecting personal medication intake in Twitter: An annotated corpus and baseline classification system.* Proc Bio Nat Lang Processing Workshop 2017, 2017. **2017**(1): p. 136-142.
131. United States Food And Drug Administration. *Approved Drug Products with Therapeutic Equivalence Evaluations (Orange Book)*. n.d.-2018; Available from: https://www.fda.gov/Drugs/InformationOnDrugs/ucm129662.htm.
132. A Kovacevic, A Dehghan, M Filannino, JA Keane, and G Nenadic, *Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives.* J Am Med Inform Assoc, 2013. **20**(5): p. 859-866.
133. A Kulanthaivel, RP Light, C Kong, K Borner, and JF Jones, *Neurological disorders and publication abstracts follow elements of social network patterns when indexed using ontology tree-based key term search.* Lec Notes Comp Sci, 2014. **2014**(6): p. 278-288.
134. DW Seo and SY Shin, *Methods using social media and search queries to predict infectious disease outbreaks.* Healthcare Informatics Research, 2017. **23**(4): p. 343-348.
135. P GUo, Q Zhang, Y Chen, J Xiao, J He, Y Zhang, L Wang, T Liu, and W Ma, *An ensemble forecast model of dengue in Guanzhou, China using climate and social media surveillance data.* Science of the Total Environment, 2018(2018 Aug 4): p. 752-762.
136. C Beisel, C Weiler-Normann, A Teufel, and AW Lohse, *Association of autoimmune hepatitis and systemic lupus erythematodes: A case series and review of the literature.* World Journal of Gastroenterology, 2014. **20**(35): p. 12662-12667.
137. J Platt, *Sequential minimal optimization: A fast algorithm for training support vector machines*. 1998, Microsoft Corporation: Redmond, WA, USA.
138. M Welling, *Fisher Linear Discriminant Analysis*. 2009, University of Toronto: Toronto, ON, Canada.
139. Y Freund and RE Schapire, *Large margin classification using the perceptron algorithm*, in *11th Annual Conference on Computational Learning Theory*. 1998: New York, NY, USA. p. 209-217.

140. G Cybenko, *Approximation by superpositions of a sigmoidal function.* Mathematics of Control, Signals, and Systems, 1989. **2**(4): p. 303-314.
141. A Kulanthaivel, CS Lammert, S Milojevic, DJ Wild, and JF Jones, *The give and take: Automated classification of support exchanged over an AIH Facebook group.* J Med Internet Res, 2019.
142. M Corrigan, GM Hirschfield, YH Oo, and DH Adams, *Autoimmune hepatitis: An approach to disease understanding and management.* British Medical Bulletin, 2015. **114**(1): p. 181-191.

CURRICULUM VITAE

Anand Kulanthaivel

Current Employment (2018-Present)
Product Content Engineer (SIFT), Clinical Architecture, LLC (Carmel IN)

Teaching Experience (2017-Present)
(All as Adjunct Faculty at the Indiana University School of Informatics & Computing, Indianapolis IN)

- **Spring 2017 & Spring 2018:** INFO B435 Clinical Information Systems
- **Fall 2017 & Fall 2018:** HIM M110 Computer Concepts for Health Information

Education

2013-2019: Indiana University School of Informatics & Computing (Indianapolis IN)
Degree: Doctor of Philosophy (Ph.D.) in Informatics
Track: Health & Biomedical Informatics
Minor: Information Architecture (Individualized)
GPA: 3.76/4.00

2011-2014: Indiana University School of Informatics & Computing (Bloomington IN)
Degree: Master of Information Science (M.I.S.)
GPA: 3.85/4.00

2000-2004: Case Western Reserve University (Cleveland OH)
Degree: Bachelor of Science (B.S.) in Biology, Molecular/Cell Track
Degree: Bachelor of Arts (B.A.) with a Major in German
GPA: 3.75/4.00

Mentorship & Degree Project Committee Membership

M.S. in Health Informatics Candidates Project Committees (numbers of students graduated)

- Spring 2018: 2
- Summer/Fall 2018: 2
- Spring 2018: 4
- Summer/Fall 2017: 4

Bibliography

Kulanthaivel A, Fogel R, Jones JF, Lammert CS. (2017). Digital cohorts within the social mediome: An approach to circumvent conventional research challenges? *Clinical Gastroenterology & Hepatology.* 15(5): 614-618.

Kulanthaivel A, Light RP, Kong C, Borner K, Jones JF. (2014). Neurological Disorders and Publication Abstracts Follow Elements of Social Network Patterns when Indexed Using Ontology Tree-Based Key Term Search. *Lec Notes Comp Sci.* 8515: 278-288.

Jones JF, Hosseini M, **Kulanthaivel A,** Pradhan M, Hosseini M. (2018). Novel approach to cluster patient generated data into actionable topics. *J Med Internet Res.* 20(7).

Holden RJ, **Kulanthaivel A,** Purkayastha S, Goggins KM, Kripalani S. (2017). Know thy eHealth user: Development of biopsychosocial personas from a study of older adults with heart failure. *Intl J Med Informatics*. 108(12): 158-167.

Jones JF, Zhang E, **Kulanthaivel A**, Katta S. (2017). What do they mean by *Health Informatics*? An analysis of employer needs vs. current curriculum competencies. *MEDINFO Proc 2017*.

Holden RJ, Voida S, **Kulanthaivel A**, Jones JF, Savoy A, et al (2016). Human Factors and User Centered Design (Book Chapter). In Clinical Informatics Study Guide: Text & Review, 1st Edition (ed Finnell J & Dixon B) Publisher: Springer

Kulanthaivel A, Lammert CS, Jones JF. (2018). A novel approach using social media to investigate patient-centric data in autoimmune hepatitis. (Poster Abstract). *Gastroenterology* 154(6): S1214.

Kulanthaivel A, Lammert CS, Jones JF. (2018). A novel approach using social media to investigate patient-centric data in autoimmune hepatitis. (Poster). Washington, DC, USA: Digestive Diseases Week 2018.

Kulanthaivel A, Patel JS, Phalakornkule K, Zhang E. (2017). The new opium of the masses: Tracking opioid use in Indiana. (Presentation). Indianapolis, IN, USA: Indy Big Data Conference 2017.

Kulanthaivel A, Kshirsagar M, Alarifi M, Oklak M. (2017). Almost Zero Error Basepair-based Record Alert (AZEBRA): A genomic clinical decision support tool. (Poster). San Francisco, CA, USA: American Med Inform. Assoc. (AMIA) Joint Summits 2017.

Kulanthaivel A, Zhang E, Katta S, Jones JF. (2016). Towards the creation of a novel career-based Health Informatics (HI) curriculum assessment instrument: Mapping HI job competencies to HI curriculum competencies. (Poster). Chicago, IL, USA: American Med Inform. Assoc (AMIA) Annual Symposium 2016.